A Systematic Literature Review of Game-based Assessment Studies: Trends and Challenges

Manuel J. Gomez, José A. Ruipérez-Valiente *Senior Member, IEEE* and Félix J. García Clemente

***Abstract*—Technology has become an essential part of our everyday life, and its use in educational environments keeps growing. In addition, games are one of the most popular activities across cultures and ages, and there is ample evidence that supports the benefits of using games for assessment. This field is commonly known as game-based assessment (GBA), which refers to the use of games to assess learners' competencies, skills, or knowledge. This paper analyzes the current status of the GBA field by performing the first systematic literature review on empirical GBA studies. It is based on 65 research papers that used digital GBAs to determine: (1) the context where the study has been applied; (2) the primary purpose; (3) the domain of the game used; (4) game/tool availability; (5) the size of the data sample; (6) the computational methods and algorithms applied; (7) the targeted stakeholders of the study; and (8) what limitations and challenges are reported by authors. Based on the categories established and our analysis, the findings suggest that GBAs are mainly used in K-16 education and for assessment purposes, and that most GBAs focus on assessing STEM content, and cognitive and soft skills. Furthermore, the current limitations indicate that future GBA research would benefit from the use of bigger data samples and more specialized algorithms. Based on our results, we discuss current trends in the field and open challenges (including replication and validation problems), providing recommendations for the future research agenda of the GBA field.**

***Index Terms*—Game-based assessment, educational technology, game-based learning**

I. Introduction

TECHNOLOGY is progressively changing the world in which we live. During the last decade, it has started to make a significant impact on educational environments, and increasing evidence has been accumulated showing the positive impact of technology in education [1]. One of the most prominent examples of technology in education is the use of digital games [2]. This type of games has become a significant part of families and, especially, among young people around the world. In fact, three-quarters of all U.S. households have at least one person who plays video games [3], while in Europe, 51% of the population aged 6-61 years play video games (an average of 8.6 hours/week) [4]. Moreover, many educators see digital games as powerfully motivating digital environments because of their potential to enhance student engagement and motivation in learning [5]. This increasing interest provides an opportunity to use video games as a tool to improve learning and education. Specifically, there is much enthusiasm in the field of education about game-based assessment (GBA) because conventional assessment methods do not seem to fully have the power to measure all aspects of students' knowledge, skills, and attributes [6].

Accompanying this explosion in technology use is the quantity, range and scale of data that can be collected, which have increased exponentially over the last decade [7]. In education, the increase in e-learning resources, educational software like Google Classroom or Kahoot, and the use of the Internet have created large repositories that provide a goldmine of educational data that can be explored and used to understand how students learn [8]. Regarding games, they allow recreating more authentic situations compared to traditional classroom activities, such as lectures or written exercises. From these situations, we can collect a vast amount of detailed data on students' interaction with the game, which provides a great opportunity to make game-based assessments (GBAs) in ways that are not possible in traditional testing [9].

In the past 10 years, numerous studies (see the work in [10] for a meta-analysis) have reported that games can be more effective for learning than other traditional teaching methods. In addition, when measuring the competencies acquired, most traditional tests present individual and decontextualized items to learners, while 21st-century competencies benefit from being applied in context for more accurate measurements. Furthermore, classic assessment often interrupts the learning process, and it does little to motivate learners [11]. Since digital games often employ challenging, interesting, and complex problems, they can be used to generate evidence of 21st-century competencies, which are traditionally difficult to measure using conventional forms of assessment [12]. The advantages of using games as a form of assessment are manifold [11], [13], [14]: they are engaging and motivating (which provides more valid assessments), and they allow us to create more complex and authentic scenarios required to assess the application of knowledge and skills. Moreover, immediate feedback based on learners' activity can reveal teachers' specific areas of difficulty to make learners keep up with the pace of the class, and such assessment would result in an adaptive game environment, which changes with learners' activity.

The implementation of assessment features into game environments is only in its early stages because it adds a very time-consuming step to the design process [15]. This situation calls for a review of the current state of the art in the GBA field for effective implementations. In this respect, we found some previous works that performed meta-reviews of the existing research on the different applications of games in learning and education. For example, the authors in [16] reviewed

Faculty of Computer Science, University of Murcia, Murcia (Spain)
Corresponding author e-mail: manueljesus.gomezm@um.es.
Manuscript received XXXXX XX, 2021; revised XXXXX XX, 2022.

137 papers to determine what empirical evidence existed concerning the effects of Game-based Learning (GBL) on 21st-century competencies and identified successful game-design elements that aligned well with established learning theories. Moreover, Alonso-Fernandez et al. [17] carried out a review focused on data science applications to game learning analytics data, showing that the primary purpose when analyzing data from serious games was assessment. Furthermore, Gris and Bengtson [18] aimed to answer how learning, engagement, and usability of games are evaluated in GBL research. To this aim, they conducted a systematic review of 91 empirical studies and categorized their measures and instruments. The researchers concluded that future research in GBL studies should add direct assessments and indirect measures to assess engagement and usability. Guan et al. [19] provided a systematic review of 35 experimental studies that substantially integrated gaming elements in primary school lessons and they noted that gamification was the most frequently used game genre. Finally, Chen et al. [20] conducted a systematic review of 146 articles related to GBL in science and mathematics education. These researchers concluded that GBL is mainly used to increase learner motivation and engagement and reduce learning anxiety. They also revealed that analyzing higher-order thinking skills (e.g., problem-solving, group collaboration) is one of the main hot topics in the community.

Despite the previous reviews of the use of games in learning in education, we have not found any specific study reviewing literature about GBA. For this reason, the current paper aims to conduct the first systematic literature review on the applications of empirical GBA studies and answer some research questions based on the analysis performed to discover current trends and open challenges in this area. The results obtained will provide an overall view of the GBA field, defining its current status and potential future steps in the research in this area.

The rest of the paper is organized as follows. Section II describes the methods, including some terminology clarifications, the research questions, databases and search terms, research selection as well as review process. Section III presents the analysis and synthesis of our results. Then, we end the paper with a discussion in Section IV and conclusions in Section V.

II. Methods

We followed a standard systematic literature review methodology, using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [21] as a basis for conducting our systematic review. First of all, (1) we formulated each research question (RQ). Then, we used (2) a fixed set of queries on a pre-identified bibliographical database, and (3) a set of inclusion and exclusion criteria. Next, (4) we made a full paper review and coding process of the RQs, and, finally, we carried out (5) a synthesis and analysis. No time restrictions were set. We can see a flow diagram representing the different stages of our systematic review (following the PRISMA template [22]) in Figure 1.

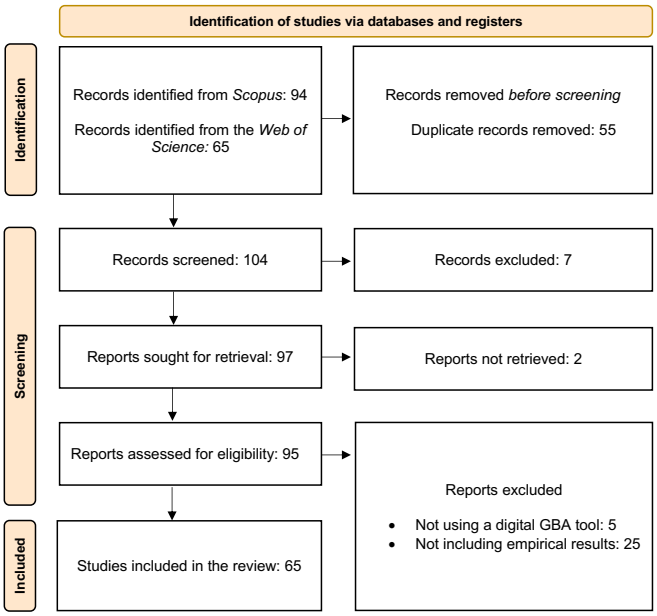

Fig. 1: Flow diagram representing the different phases of the systematic review.

A. Terminology clarifications

In this Section, we present a set of definitions that aim to clarify the concept of GBA, which is the focus of this systematic literature review. Firstly, we can define a **game** as "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome" [23, p. 80]. In addition, games also have clearly-defined goals and obstacles for the player to overcome, providing only intrinsic rewards (satisfaction for getting the right answer) [24]. Secondly, **GBL** can be regarded as an innovative learning approach where a game is developed to deliver immersive and attractive learning experiences aiming at particular learning goals, experiences and results [25]. Thus, GBL uses a game containing learning content derived from school curricula or essential life skills to improve the learning experience. Moreover, **GBA** is a specific application of games, referring to a type of assessment that uses players' interactions with the game, both digital and non-digital, as a source of evidence to make meaningful inferences about what players know and can do (i.e., knowledge, skills), and how individual players interact with the game as a problem-solving process [15], [26]. Finally, we have the concept of **gamification**, which is usually defined as "the use of game design elements in non-game contexts" [27, p. 9].

Although GBL and GBA are often confused with gamification and gamified assessment, it is undeniable that some differences exist between them. While GBL implies the use of a game developed for learning purposes, gamification utilizes game elements in non-game contexts, not necessarily using full games inside the activities [28]. Thus, GBA also implies the use of a game developed for assessment purposes, using players' interaction with the game as a way to obtain evidence and use this evidence as a form of assessment. Therefore, tools that use gamified activities to assess students' knowledge (e.g.,

Kahoot, Duolingo) use gamified assessments, and cannot be considered as GBAs. We can also make a clear distinction between GBA and a simple measurement using games since GBA is intended for evaluating players' skills or knowledge based on their interaction with the game. As Ghergulescu and Muntean state, "measurement represents the process of collecting the information needed for assessment" [29, p. 357]. In other words, measurements are used as evidence to make meaningful inferences about what players know and can do, while measurements using games do not perform that evaluation. These are the definitions that we applied as part of the systematic review screening process to consider a given paper within the GBA field or not, including or discarding that study.

B. Research questions

To state each one of the RQs, we analyzed and simplified the steps in empirical GBA research [30], which can be seen in Figure 2.

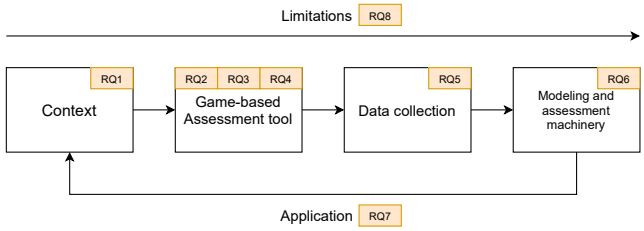

Fig. 2: A simplified view of the steps in GBA empirical research.

In this process, we can identify five different stages: (1) Learning environments, with the context and learners; (2) The GBA tool that is going to be used in the research; (3) Data collection, to identify which data has to be collected and how to store them; (4) Modeling and assessment machinery; and (5) Educational application, to identify the final objective and target users. From these stages, we identified the following RQs, which allow us to understand the open challenges and current trends in the area:

RQ1. In what context or environment has GBA been applied?
RQ2. What is the primary purpose of GBA?
RQ3. What is the domain of GBA?
RQ4. Is the game/tool used available to the public?
RQ5. What is the size of the data sample used in the study?
RQ6. What computational methods and algorithms have been applied in the research?
RQ7. What stakeholder is the intended recipient of the research results?
RQ8. What limitations and challenges do the authors address?

In addition, Figure 2 also shows the mapping between the different stages and the RQs identified: RQ1 is based on the first stage, related to the learning context. Then, RQ2, RQ3, and RQ4 are based on the second stage, which refers to the GBA tool used, its primary purpose, the domain and availability. Next, we wanted to investigate the sample size (RQ5) which is situated in the data collection stage. Regarding the modeling and assessment machinery, our objective was to investigate the computational methods and algorithms applied in the research (RQ6). RQ7 refers to the application of the research in the desired context, identifying the main stakeholder of the research. Finally, RQ8 aims to identify the research limitations at any stage.

C. Databases and search terms

We have queried two databases: Scopus and the Web of Science since they are the most widely used databases in different scientific fields and are often used for surveying the literature [31]. Scopus is the world's largest citation database of peer-reviewed research literature, with over 22,000 titles (including journals, conferences and book series) from more than 5,000 international publishers, of which 20,000 are peer-reviewed journals in the scientific, technical, medical, and social sciences [32]. Moreover, the Web of Science, the second biggest bibliographic database, can be used to track ideas going back several decades from almost 1.9 billion cited references from over 171 million records [33].

To perform the search on both databases, we restricted the query to title and keywords: 1) we included the term "game-based assessment" and searched for it within the paper titles; 2) we included the term "game-based assessment" and searched for it within the paper keywords. Thus, we used the following final search query:

(TITLE("game-based assessment") OR *KEY*("game-based assessment")).

The initial selection of studies was retrieved in January 2021, and this query generated 159 initial studies (94 from Scopus and 65 from the Web of Science).

D. Inclusion/Exclusion criteria

After obtaining the initial collection, we excluded the duplicated records from the two databases (55 studies). Then, we made a first brief review of all papers, comparing them against the inclusion and exclusion criteria. This first review was conducted by one of the authors. After the first analyses, we classified studies as *included* or *excluded*, and the coding results were discussed collaboratively by the three authors in order to obtain the final set of included and excluded studies and avoid possible errors. The inclusion/exclusion criteria followed are described below. Given these criteria, the paper was included if all of the following conditions were met (i.e., if one condition was not met, the paper was excluded). Furthermore, the conditions were applied sequentially, so that a paper not matching a condition was excluded immediately from the collection:

- The paper was written in English or Spanish (languages in which the authors have high proficiency): 0% of the papers were excluded.
- The paper was fully accessible: 1.9% of the papers were excluded (2 studies).
- The paper was published in conference proceedings, journals or edited books/volumes (i.e., book chapter): 0% of the papers were excluded.

- The paper was not extended at a later time (i.e., a conference paper that was later on extended in a journal paper): 6.9% of the papers were excluded (7 studies).
- The paper used a digital GBA tool: 5.3% of the papers were excluded (5 studies). See Section II-A above for relevant definitions.
- The paper included empirical evidence related to the outcomes of applying the GBA tool: 27.8% of the papers were excluded (25 studies).

E. Final paper collection

After the first brief review to ensure that every paper met our inclusion/exclusion criteria, we excluded a total of 39 papers. Thus, the final paper collection consists of 65 studies.

Figure 3 shows the distribution of papers within the final collection by publication year. We see an increasing interest in this particular topic: between the years 2013 and 2016, we only have 21 (32.3%) published papers that matched our criteria, while between 2017 and 2020 there are 44 (67.7%) of them.

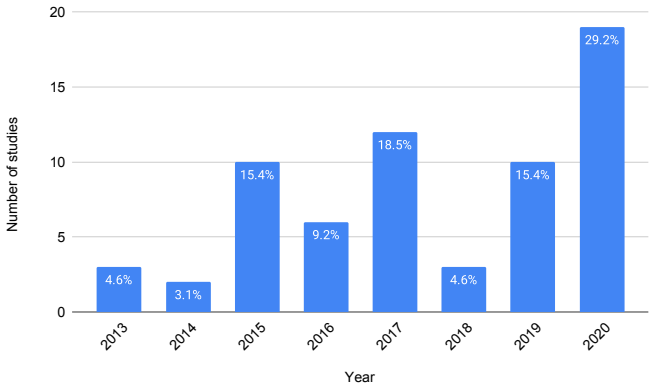

Fig. 3: Number of selected studies and rate in the collection per year of publication.

We also collected each paper's keywords and made a brief analysis to describe our paper collection. For our analysis, we excluded the "game-based assessment" keyword since it was the most common one. The most frequent keywords are presented in Figure 4. The total sum of keywords is 299 while there are 201 unique keywords. The average keyword was found 1.48 times. As we can see, the predominant keywords strongly focused on games, assessment, and analytics.

F. Review and coding process

In the coding stage, we collected the data of the selected studies that we consider to be the most valuable to address the RQs in Section II-B. Based on the aim of the review, we followed an inductive coding scheme (also called open coding). This means that the codes created were based on the qualitative data itself [34]. This is an iterative process since researchers can add new codes, split an existing code into two, or compress two existing codes into one as they continue reviewing data. Specifically, in our analysis, we first made a brief review of each paper (conducted by one author), collecting all the necessary information to code each RQ at once. After that, we followed an iterative process whereby we continued reviewing the information corresponding to each RQ sequentially, and unclear results were discussed and contrasted by the three authors. The full results of the coding process per paper are available in [35]. In addition, it should be noted that each paper can fit into more than one of the codes created for each RQ.

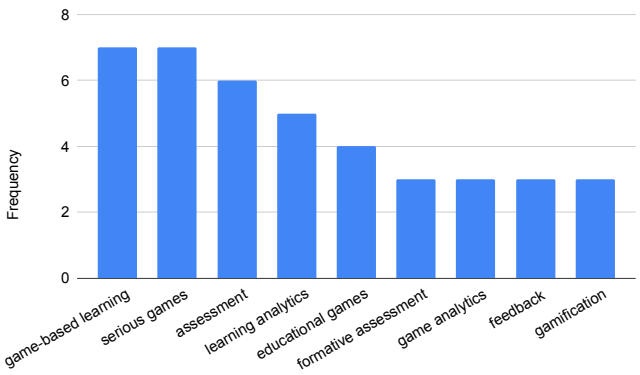

Fig. 4: Distribution of keywords across articles in the final collection.

III. Results

A. In what context or environment has the GBA been applied? (RQ1)

GBAs can be used in very different environments. Our analysis reveals that there are three main contexts where GBAs have been used:

1) `K-16 education`: some papers use GBAs in K-16 education (e.g., school, university) to support teaching and learning. More specifically, games are most commonly used in middle school and high school (23.1%). However, games are also used in other K-16 education environments such as primary school (15.4%) and university (10.8%). For example, Di Cerbo et al. [36] used game data from 751 US middle school players.
2) `Medical`: games can also be used in medical environments for different purposes (e.g., rehabilitation). For example, the authors in [37] examined the feasibility of administering the GBA in a sample of inpatients with chronic schizophrenia with low levels of functioning. Moreover, the authors in [38] aimed to present data on construct validity, test–retest reliability and feasibility, measuring motor-cognitive functions in multimorbid patients with mild-to-moderate dementia. Regarding construct validity, the authors tested eight hypotheses and confirmed seven of them (87.5%), thus indicating excellent construct validity. Moreover, the authors found moderate-to-high test–retest reliability (ICC=0.47-0.92).
3) `Workforce`: another option is the use of games to assess in professional environments. In this context, enterprises can use games to evaluate their employees or provide them with additional feedback. Even now,

companies can include the use of GBA for the recruitment of staff and the selection process [39]. This idea is supported by the fact that in-game constructs show similar relationships with in-game performance to what the workforce constructs do with job performance [40].

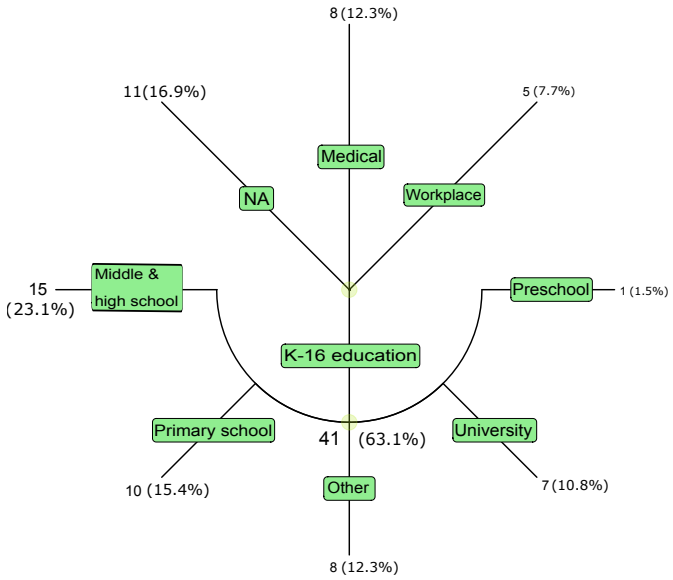

Fig. 5: Papers' category distribution based on RQ1.

There are also some studies such as [41] that did not specify in what context their games were used (11 studies, 16.9%). We can see the number of papers fitting in each category in Figure 5. As the figure shows, GBAs are mostly used in `K-16 education` (41 studies, 63.1%), followed by `medical` (8 studies, 12.3%) and `workforce` contexts (5 studies, 7.7%).

B. What is the primary purpose of the GBA research? (RQ2)

In this RQ, we wanted to know what was the main purpose of each GBA in each study. We coded the papers' main purpose into six different categories: `GBA evaluation`, `study of in-game behaviors`, `assessment`, `interventions`, `framework proposal` and `game design proposal`. Next, we describe in detail each one of these categories.

1) `GBA evaluation`: in these studies, authors evaluate the game by checking if it achieves its initial objectives using some measure to prove that the game or tool is suitable for an educational environment. In [42], the authors showed how they applied the methodology for an assessment game for ICT managers in secondary vocational education, checking if this assessment was content-valid compared to a face-to-face assessment. Moreover, the authors in [43] aimed to investigate whether it is possible to perform an in-Basket test (which is widely used by companies and organizations in order to map employees' soft skills) online with the same effect as that of the onsite one.
2) `In-game behaviors`: in these studies, authors investigate in-game players' behaviors (e.g., persistence, engagement). By identifying these behaviors, we can group players according to different behaviors or simply check if a student shows a specific one. For example, Dicerbo [11] used evidence extracted from log files to create a measure of persistence. Similarly, Ventura & Shute [44] also created a measure of persistence, validating it against another existing measure and concluding that the GBA predicted students' learning.
3) `Assessment`: in these studies, games are used to report measures that aim to evaluate students. This allows for improvements in the learning process using this evaluation measure instead of classic evaluation methods or providing personalized feedback. In their work, Weiner & Sanchez [45] created an alternative measure using a virtual reality game that calculated scores to indicate specific cognitive abilities.
4) `Interventions`: games can also be used to investigate the effect of some interventions while playing. For example, we can use feedback messages to notify the learner with positive (or negative) feedback to observe how this intervention influences its performance and behavior. Another typical example is switching the order of in-game elements or testing different game features. In [46], the authors used a psychophysiological methodology to investigate attention allocation to different feedback valences (i.e., positive and negative feedback). With that purpose in mind, they used an eye tracker to collect accurate information about individuals' locus of attention when they process feedback.
5) `Framework proposal`: in these papers, the authors propose the design of a new framework to be used within the context of GBA. We can see an example in [47], where the authors examined the process of creating a Bayesian network framework through different techniques (e.g., using correlation matrixes, IRT) to create scoring models for assessing students.
6) `Game design proposal`: authors provide a game design that can be used for assessment purposes. For example, the authors in [48] show the design of an online GBA to help students improve their learning outcomes and promote the development of general and transferable skills, such as the ability to solve problems in complex situations, and working under pressure.

Some studies focused on more than one of the categories described above. For example, Weiner & Sanchez [45] used a virtual reality game to calculate a score measure for each student (`assessment`) and they proved that these calculated scores are best used by comparing them to classic measures (`GBA evaluation`).

We can see the number of papers fitting each category in Figure 6. `GBA evaluation` is the most common category (38 studies, 58.5%), followed by `assessment` (34 studies, 52.3%) and `framework proposal` (12 studies, 18.5%). The less common category is `game design proposal`, with only three papers fitting (4.6%). We can conclude that most papers focused on using games to assess learning, but they also tried to prove that this assessment was a valid

measure to be used in real educational contexts.

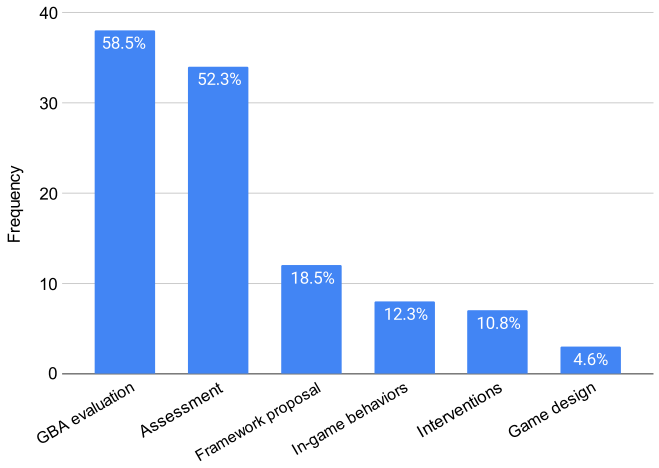

Fig. 6: Number of studies and rate in the collection per purpose of the research. More than one purpose is possible for a given paper.

C. What is the domain of the GBA? (RQ3)

From reviewing the selected papers, we identify four major domain categories: `STEM`, `humanities and social sciences`, `cognitive and soft skills`, and `physiological capacities`. As some of the categories have more than one related area, we also consider some sub-categories fitting into them. We describe each domain category below in detail:

1) `STEM`: in this category, we include papers that are related to `science`, `technology`, `engineering` and `mathematics`. For example, Chiu & Hsieh [49] showed the different teaching methods of second-grade elementary students in fraction concepts (`mathematics`), while Kim et al. [26] aimed to assess the understanding of Newton's three laws of physics using a two-dimensional physics game.
2) `Humanities and social sciences`: papers related to humanities and social science areas (e.g., art, music, language) fit in this category. As this is a wide area, we have also defined some sub-categories to better categorize the papers. These sub-categories are `language`, `art` and `history`. Studies that do not fit into one of those three categories are categorized as `other`. As an example of the `art` category, we highlight the work in [50], where the authors used a game in which players collect data about the musical interests of an in-game character and use these data to make decisions about which artists to sign and what songs to record. We can see another example (related to `language`) in [51], where the researchers described the design of an argumentative reasoning task within a scenario-based assessment enhanced with game elements.
3) `Cognitive and soft skills`: cognitive skills are the core skills your brain uses to think, read, learn, remember, reason, and pay attention [52]. Cognitive skills help to process new information by taking that information and distributing it into the appropriate areas in the brain. Developing cognitive skills helps to complete this process more quickly and efficiently, helping people to understand and effectively process new information [53]. Moreover, soft skills are described as a combination of interpersonal and social skills, including the ability to communicate, coordinate, work under pressure, and solve problems [54]. In this category, we consider `attention`, `memory`, `logic and reasoning`, `visual processing and speed`, and `soft skills`. We find papers that have measured interesting skills, such as [55], which included a series of reasoning activities to measure argumentation skills (which is related to `logic and reasoning`), or [56], [57], which used GBAs to assess candidates' soft skills.
4) `Physiological capacities`: physiological functional capacity is the ability to perform the physical tasks of daily life and the ease with which these tasks can be performed. We could assess daily physical tasks, like Rodríguez de Pablo et al. [58], who used a set of games to provide a fast, quantitative and automatic evaluation of the arm movement function. Furthermore, other works focused on assessing mental abilities, such as motivating children with autism to make more eye contact [59].

There are also papers fitting more than one category at once. For example, in [60], [61], researchers used a GBA for measuring argumentation and pragmatic skills. This research measured language competencies (which is part of `humanities and social sciences`), but it also measured `cognitive and soft skills`. We can see the full tree showing the distribution of studies into categories and sub-categories in Figure 7.

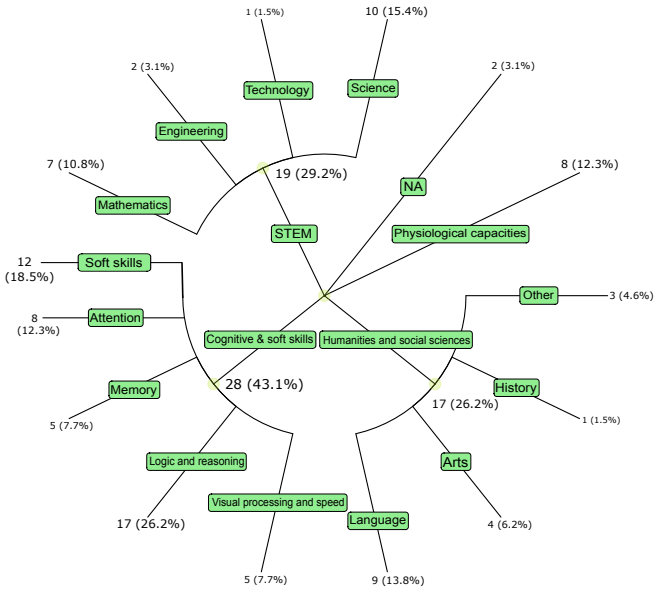

Fig. 7: Category tree for RQ3.

The three predominant categories are `cognitive and soft skills` (28 studies, 43.1%), `STEM` (19 studies, 29.2%), and `humanities and social sciences` (17 studies, 26.2%). Taking a look at each sub-category, we note that the main area in `STEM` is `science` (10 studies, 15.4%). The main field in `cognitive and soft skills` is `logic and reasoning`, with 17 papers (26.2%), while in `humanities and social sciences`, the predominant sub-category is `language`, with nine papers (13.8%).

D. Is the game/tool used available to the public? (RQ4)

A critical aspect of research is the availability of the results obtained to be used by the general public. It is essential to make tools accessible so that researchers can replicate experiments and practitioners can use them as part of their teaching. From our analysis, we find three primary categories: `Currently available`, `Not available (NA)` and `Not specified`.

1) `Currently available`: the game/tool used in the corresponding research is currently available (using the web portal specified by the authors) for public use (e.g., [62], [63]).
2) `Not available`: the game/tool used in the research was presented as initially available in the paper, but currently, it is no longer accessible based on our attempt to access the site (e.g., [64]–[66]).
3) `Not specified`: researchers did not specify the tool's availability; it is more than likely that it is not accessible (e.g., [67]–[69]).

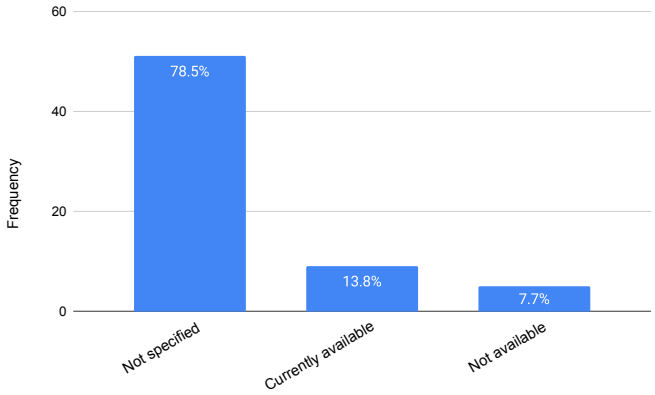

Fig. 8: Number of studies and rate in the collection per research availability.

Figure 8 shows that most papers (51 studies, 78.5%) did not specify if the tool is accessible or not. Another minority of papers (9 studies, 13.8%) offered their tool publicly. The rest of the studies (five papers, 7.7%) initially offered their games, but they are currently unavailable. In addition, we did not find any open-source game across the studies included in the review.

E. What is the size of the data sample used in the study? (RQ5)

In this research question, we classify the different data collections used in the studies based on their data sample size. Investigating the sample size is relevant since the use of larger data samples will allow better generalization of the research results, as well as the possibility of applying more complex algorithms (e.g., neural networks), which often require large amounts of data to outperform other models [17]. Although the sample size can be relevant for some aspects, such as preventing overfitting in some methods, it is not related to the study's rigor (i.e., using a larger sample does not make a study more rigorous). From the coding process, we present four categories:

1) `Fewer than 50 participants`: these papers involved fewer than 50 participants in their empirical studies. We find studies with small data samples, such as [70], using data from 30 postgraduate students, or [71], which used a sample of 20 healthy controls patients and 18 patients with Alzheimer's disease to evaluate the usability of a tool created to assess cognitive functions.
2) `Between 50 and 250 participants`: these papers involved between 50 and 250 participants in their studies. For example, Leonardou et al. [72] used data from 77 primary school pupils for assessing and improving multiplication skills. We see another example in [73], which used data from 95 children from the final year in preschool to measure psychoacoustic thresholds.
3) `Between 250 and 500 participants`: these papers involved between 250 and 500 participants in their studies. For example, Gjicali et al. [74] used data from 433 students who played a game simulating an artificial culture with norms embodying two cultural concepts: hierarchy and collectivism.
4) `More than 500 participants`: these papers used data from more than 500 participants in their research. Hautala et al. [75] used data from 723 students to investigate reading difficulties, concluding that the GBA could be successfully used to identify students with reading difficulties with acceptable reliability (Cronbach's alpha 0.93 and 0.87). Some other studies used a huge sample, such as [12], which used data from 5,545 students to measure engagement and cluster students to finally report four different engagement profiles.

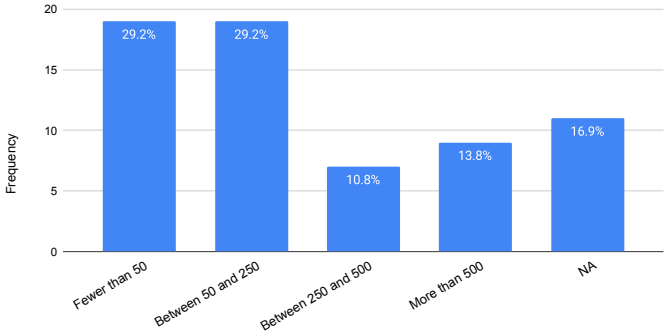

Fig. 9: Number of studies and rate in the collection per data sample size.

Other papers (e.g., [76], [77]) did not specify the data sample size of the study and we categorize them as `NA`. Figure

9 summarizes the results of the data sizes across papers. We can see that only 16 papers (24.6%) used more than 250 participants in their students, and only nine papers (13.8%) used data from more than 500 students, meanwhile most of the papers (58.5%) used data from fewer than 250 participants. We also see that a significant amount of papers (11 studies, 16.9%) did not specify the data sample size in their studies.

F. What computational methods and algorithms have been applied in the research? (RQ6)

After exploring the data samples that were retrieved across papers, our goal was to examine the methods that were applied for its analysis. We believe that the methods being applied are crucial, since they are the link between the evidence generated by learners and the assessment. Accordingly, we identified five different groups of methods for analyzing the data: `Descriptive statistics`, `Machine learning`, `Knowledge inference`, `Deep learning`, and `Sequence mining`. Below is a description of each group in detail:

1) `Descriptive statistics`: these encompass further mathematical analyses covering various methods, tests, and visualizations. We identified several papers that applied `summary statistics` (e.g., mean, variances) [78], `correlations` [79] and `visualizations` [80].
2) `Machine learning`: it is a part of AI and covers a set of methods that allow systems to learn and improve from historical data automatically. We noted that the authors used two significant families of machine learning methods: `supervised learning` and `unsupervised learning`. `Supervised learning` includes techniques such as regression [81] while `unsupervised learning` uses other methods, such as clustering techniques like *k*-means [12] or dimensionality reduction techniques like Principal Component Analysis (PCA) [82].
   For example, the authors in [83] developed a game for evaluating the logic abilities of first-year university students. They tried to compare the measures obtained by paper-based tests with those obtained using the game by conducting a linear regression (which is a supervised method). The authors concluded that the measures obtained from both methods were not significantly different.
3) `Knowledge inference`: it refers to the acquisition of new knowledge from existing facts based on certain rules and constraints. One way of representing these rules and constraints is through the use of logic rules, formally known as knowledge representation [84]. Common knowledge inference methods that several studies have used are `Bayesian networks` [85] and `fuzzy cognitive maps` [67].
   In [86], the researchers proposed a dynamic Bayesian network modeling approach for measuring student performance from an educational video game. The results supported the usefulness of Bayesian networks to characterize and accumulate evidence regarding students in games and related assessment environments.
4) `Deep learning`: an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for decision-making [87]. An example is the work of Chen et al. [88], who used Long Short-Term Memory (LSTM), an artificial recurrent neural network architecture.
5) `Sequence mining`: the objective of sequence mining is to unlock useful knowledge hidden in sequence data [89]. Specifically, Gomez et al. used [90] sequence mining to identify sequences and errors by transforming raw data into meaningful sequences that are interpretable and actionable for teachers.

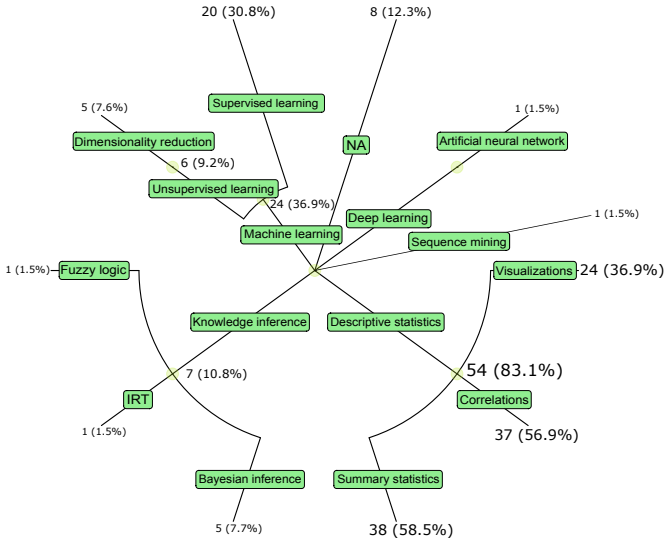

Fig. 10: Paper distribution based on the methods used.

In Figure 10 we can see the different families of techniques and the number of papers that used them in their research. We can see that most papers (83.1%) used `descriptive statistics`, and almost none of them used `deep learning` (only one paper, 1.5%). We also noted that 83.3% of the papers that used `machine learning` techniques used `supervised learning` too, specifically, most of them used different types of regressions.

G. What stakeholder is the intended recipient of the research results? (RQ7)

A stakeholder is defined as a person with an interest or concern in something, especially a business [91]. In our study, we consider the paper's stakeholder as the person to whom the results are directed, even though the paper's contribution might have other secondary stakeholders. Specifically, we have two main groups of stakeholders: `researchers` and the `final user`.

1) `Researchers`: if the paper's contribution is methodological, we expect that the paper's main stakeholders will be researchers. For example, Lonergan et al. [92] created a Paper-based Assessment (PBA) and a GBA in

order to measure students' performance, cognitive states and satisfaction related to both assessment methods. The authors concluded that smaller versatile GBAs may have a greater impact on the student's cognitive capabilities, and could enhance student performances during, for example, a final exam or short formative assessments. Moreover, Tsai et al. [93] proposed an online learning system using different gaming modes of classic tic-tac-toe games to explore how different gaming modes and feedback types in this game-based formative assessment affect knowledge acquisition effectiveness and perceptions of participation.

2) `Final user`: if the paper's results are to be used by final users or are validating the GBA, we consider that the main stakeholder will be the final user in that context (e.g., teachers and students). In their work, Ciman et al. [94] designed a game to support children with cerebral visual impairment, developing a mobile version of the game to be used by children easily at home on any platform. Delcker & Ifenthaler [95] also developed a mobile app that makes an automated analysis of the data and provides information about children's language skills. Other papers focused on teachers, such as [96], where the authors used a GBA to develop a set of visualizations to support teachers in classrooms.

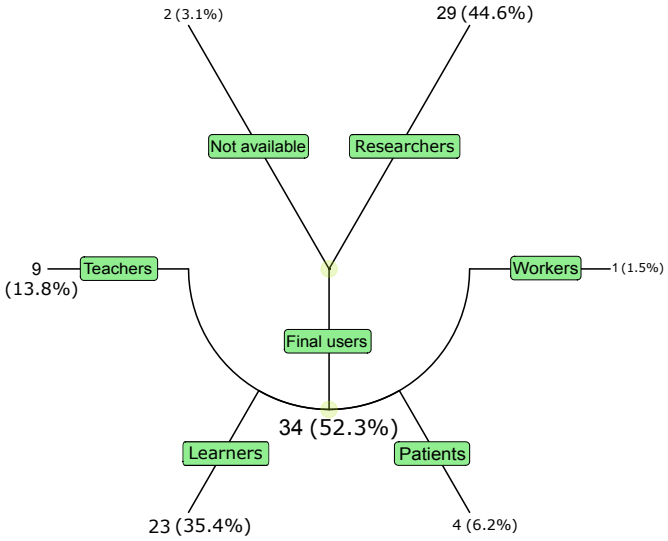

Fig. 11: Paper distribution tree based on the main stakeholder.

Figure 11 shows the number of studies that focused their work on each of the stakeholders. We see that 34 studies (52.3%) were directed to `final users`, mainly students. Moreover, 29 studies (44.6%) focused on `researchers` as the main result recipients. Two further studies (e.g., [97]) did not provide results and we categorized them as `NA`. Focusing on studies directed to the final user, we see that the majority of papers are directed to `learners` (35.4%) and `teachers` (13.8%).

H. What limitations and challenges do the authors address? (RQ8)

Limitations show potential weak points of the study that researchers usually highlight regarding their work such as constraints in research design or methodology. We can group the limitations that the authors faced in the six following categories: `game design`, `data sample`, `methodological`, `technical`, `integration` and `validation`.

1) `Game design`: an appropriate design of a game is crucial for learners' assessment since the GBA design must be adapted based on the constructs that will be evaluated. It requires a great effort to design a good GBA, aligning the evidence collected with the final purpose of the assessment. Designers might have different goals when developing a GBA [14]: "for game design, engagement; for instructional design, developing key concepts and capabilities in the target domain; for assessment design, evoking evidence of those capabilities for the intended use case(s)." Moreover, many game design decisions play a role in what kind of game performance is achieved and its meaning [98]. Future designers should consider these concerns to achieve better designs, thus, creating more engagement and facilitating the development of the key concepts and capabilities intended for learning.
2) `Data sample`: data were crucial for our review because GBA is based on the evidence, stored as data, generated by the students' interaction with the games. We examined each paper and found several limitations related to data. Jackson et al. [99] reported that they had a small sample size and that larger sample sizes would be necessary to detect smaller effects. We see a similar example in [100], where the authors had a sample collection of 67 students, but only four of those 67 student samples were used in their empirical study. The work in [101] described the difficulty of designing a good data model, as there are usually conflicts between programmers and assessment designers, usually complicated by constraints related to budgets and schedules.
3) `Methodological`: this category includes challenges and limitations related to the methods, algorithms, or techniques used. For example, Yu et al. [102] wanted to collect additional data to explore learners' behavioral patterns during gameplay. We see another example in [103] since the authors reported that the assessment developed in this study only includes a part of number sense (this term refers to a group of key math abilities), and, in order to complete the number-sense battery, the assessment tools for the other components of number sense are needed to be developed.
4) `Technical`: it is defined as a challenge involving how a machine or system works. This could include storage limitations, computing power, or even limitations related to sensors used in the study. In [104], the authors pointed out the necessity of a database (to store information about students' achievements), since they could not store that information, as well as the necessity of an admin-

istrator module to facilitate developing and modifying game elements.

5) `Integration`: incorporating game activities as part of the curriculum in schools remains limited due to certain factors such as the schools' budget or the rigidity of subjects' classic curriculum. Halverson & Owen [64] claimed that if GBAs can show that games can serve as assessments that generate reliable evidence, we could then legitimize the potential of games and then break the social conventions that limit the potential of learning and assessment technologies in schools.
6) `Validation`: one of the most significant parts of the research is the validation of the results. Validation is intended to ensure that the proposed methods and the accomplished results proved satisfactory by conducting empirical experiments. Sanchez & Langer [105] suggested that the games used in their study were entertainment games, and further research could be oriented to validate their results with games designed for assessment purposes.

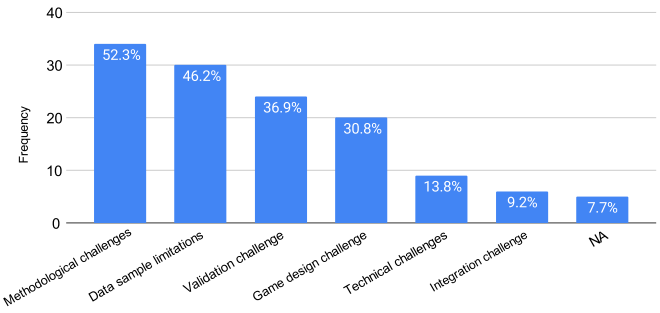

Fig. 12: Number of papers based on their limitations and challenges.

Some other studies did not report any challenges or limitations (e.g., [106]). Figure 12 shows that `methodological` challenges are the most common ones (34 studies, 52.3%), followed by `data sample` limitations (30 studies, 46.2%). On the other hand, the least frequent limitations are related to `integration` (6 studies, 9.2%) and `technical challenges` (9 studies, 13.8%).

IV. Discussion

In this section, we first present a summary and discussion of our main findings. We can also see a summary of these main findings in Table I. Next, we present a discussion of past and future challenges regarding games for assessment. Finally, we address existing limitations in this study and the implications of our research.

A. Current trends

First of all, we analyzed the contexts where the studies were applied (RQ1), finding that most of them took place in `K-16 education`, especially in `high school and middle school`. This is an exciting finding because young kids and teenagers represent the major force whose 21st-century competencies development will be heavily impacted by technology [16], [107]. Moreover, children and adolescents are an ideal target since the familiarity of these users with gaming environments and game mechanics facilitates their interactions with games [17].

Research question	Categories	Count	%
Context (RQ1)	Formal education	41	63.1%
	Medical	8	12.3%
	Workplace	5	7.7%
	Not Available	11	16.9%
Main Purpose (RQ2)	GBA evaluation	38	58.5%
	In-game behaviors	8	12.3%
	Assessment	34	52.3%
	Interventions	7	10.8%
	Framework proposal	12	18.5%
	Game design proposal	3	4.6%
Domain (RQ3)	STEM	19	29.2%
	Cognitive and soft skills	28	43.1%
	Humanities and social sciences	17	26.2%
	Physiological capacities	8	12.3%
	Not Available	2	3.1%
Availability (RQ4)	Not specified	51	78.5%
	Currently available	9	13.8%
	Not available	5	7.7%
Sample Size (RQ5)	Fewer than 50 participants	19	29.2%
	Between 50 and 250 participants	19	29.2%
	Between 250 and 500 participants	7	10.8%
	More than 500 participants	9	13.8%
	Not available	11	16.9%
Algorithms/ techniques (RQ6)	Descriptive statistics	54	83.1%
	Machine learning	24	36.9%
	Deep learning	1	1.5%
	Sequence mining	1	1.5%
	Knowledge inference	7	10.8%
	Not available	8	12.3%
Stakeholder (RQ7)	Researchers	29	44.6%
	Final user	34	52.3%
	Not available	2	3.1%
Limitations (RQ8)	Technical	9	13.8%
	Game design	20	30.8%
	Data sample	30	46.2%
	Methodological	34	52.3%
	Integration	6	9.2%
	Validation	24	36.9%
	Not available	5	7.7%

TABLE I: Summary of the main findings.

Regarding RQ2, we found that the majority of GBA studies focused on students' `assessment` and the `validation` of the game or the tool used. This suggests that having established that games are helpful for other purposes beyond entertainment, there is an increasing interest in using games as a natural alternative to classic evaluation methods, validating and comparing them against those traditional alternatives. Moreover, the fact that researchers also focused on the validation of the GBA used is a promising finding. Specifically, Gris and Bengtson [18] pointed out the lack of evidence about engagement and usability needs, especially with well-assessed reliability and validity. We also noticed that few studies had the main purpose of proposing or validating a `game design` for assessment. Although many studies proved that GBA could improve students' learning outcomes, we should not forget game design. The literature reveals that game design is essential, and several distinctive design elements, such as narrative context, rules, goals, rewards, multi-sensory cues, and interactivity, seem necessary to stimulate the desired outcomes [108], [109].

We also extracted four predominant domains (RQ3) across studies. A large proportion of the analyzed studies aimed at practicing and assessing content related to `STEM` as well as `humanities and social sciences`. This is not surprising since many of the studies took place in schools and high schools, and the use of games in these contexts is an ideal opportunity to teach content related to the main subjects at those ages. Another large number of papers also focused on developing and measuring `cognitive and soft skills`. Using game design as a context to teach higher-order thinking skills has drawn attention from researchers since schools usually place heavy emphasis on covering and delivering content knowledge [110]. Moreover, this could be useful not only in educational contexts, as we have seen some studies that measure cognitive skills for `medical` purposes (e.g., rehabilitation) [68], [71]. However, the researchers in [111] pointed out the lack of research on 21st-century skills such as creativity and critical thinking.

We discovered that many of the studies had small data samples (RQ5). Furthermore, a significant part of the studies did not specify the data sample size used in the experiment. This is also noticed by the researchers themselves, as nearly half of the studies reported data sample limitations. Moreover, apart from collecting the sample size, we also tried to collect information about the type of data collected. However, almost no study included information related to the type or format of the data used.

Across papers, researchers used many different algorithms and techniques (RQ6) to analyze the data. We classified them into five categories and found that the most common ones are `descriptive statistics` and `machine learning`. With that said, we note that the majority of papers used statistical analyses or basic machine learning algorithms, and few studies used more complex or advanced methods, which might be more adequate to model students' knowledge properly. However, those techniques that are easier to implement are also the ones chosen more frequently by researchers. Therefore, more work is needed to develop specialized GBA methods, that are also affordable to implement. This could perhaps be done through open-source libraries and more reproducible research. Moreover, we consider that making results interpretable is an essential part of the assessment, and one way to reach this interpretability is by using `visualizations`. Visualizations are essential components of research presentation and communication because of their ability to represent large amounts of data [112] and because it is easier for the brain to comprehend an image as opposed to words or numbers [113]. We think this is a promising way to integrate games in schools, and we realized that studies now tend to use visualizations to communicate their results (e.g., [62], [80]).

Finally, we wish to report the scarce information regarding games and tools availability (RQ4). The majority of studies did not provide any information on how to access or use their tools. In addition, some studies made tools public but expired, being inaccessible at present. This underscores the low transference of this research to practice and, thus, we encourage authors to make their products and results publicly accessible since we consider that this is an essential part of this type of research.

B. Open challenges

From our results and previous related reviews, we find some open challenges in the area that authors usually report. A description of each of these challenges is found below:

In [114], the authors address the challenge of how to make appropriate assessments. They noted that pre-test and post-test measures are a good manner to make an assessment, and they also recommended unobtrusive ways to collect data, such as another person taking notes during game-play. In our review, we noted that, at present, most studies found an appropriate method to make good assessments using Evidence-centered Design (ECD) and stealth assessment. ECD framework views assessment as an evidentiary argument, that is, an argument from which we observe what students say, do, or make in a few particular circumstances [115]. Moreover, stealth assessment represents a unobtrusive, yet powerful process by which learner performance data are continuously gathered during playing and learning, and inferences are made about the level of relevant competencies, maintaining the learners' flow and engagement [116]. Since ECD and stealth assessment are two common practices in current research, we could claim that the objective of making appropriate assessments using unobtrusive methods has been accomplished.

What data are going to be collected is as important as how to collect these data, and another present challenge is the design of games for specific assessment purposes. Akcaoglu & Koehler [110] indicated that games that present a hidden questionnaire format do not engage learners, while well-designed games can engage learners in reflective thinking [117]. Although we identified a few papers with the main objective of providing a good game design, many of them have developed an excellent game for other purposes. Some examples are [12], [71], and [73]. Future research should focus on complex game designs rather than the typical simple quiz design, employing multiple game-design elements such as collaboration, role-playing, narrative, exploration and complexity [16].

An important open challenge at present is replication and transferring the research to practice. In addition to the findings in our literature review about the game or tool being unavailable in most cases, All et al. [118] mentioned replication issues with certain studies due to missing information in multiple areas of the study. It is crucial to provide a detailed description of the procedure followed to conduct the study. While the community is currently demanding more standardized open science practices, this problem is still present currently. Besides, Alonso-Fernandez et al. [17] noted that most papers did not describe the format in which they collected the data, so we cannot know if they used a standard or relied on their data formats, which represent even more replication and reusability issues. In addition, having open-source games or tools would be especially helpful for researchers. Unfortunately, we did not find any available open-source games across the studies. This problem of missing information is a familiar issue in

multiple research fields (nearly every field is affected), and it leads to other problems such as low reproducibility. In fact, the terms "reproducibility crisis" and "replication crisis" have gained significant popularity over the last decade [119]. To fix this issue, the community is demanding more pre-registered studies, open data, open analyses, and open access publications [120], and this can be systematized by the guidelines of the publishers, governments and research communities [121].

Regarding the methods and techniques, we identify the challenge of implementing learner modeling algorithms. As we mentioned above, researchers usually use simple techniques to conduct their studies. In addition to Alonso-Fernandez et al. [17] noting that limitation, we confirmed it in our results. In our review, 52.3% of the papers reported methodological challenges to be addressed in future research, most of them related to the use of more complex metrics and techniques to infer new information. It is important to benefit from more advanced techniques (e.g., knowledge inference techniques, deep learning techniques) that can allow us to infer more complex and valuable information from the data collected. However, an important limitation of many of those advanced techniques is their low interpretability. Even if visualizations are a promising way to improve the presentation of results and communication, they cannot improve the model's interpretability themselves. According to the researchers in [122], with machine learning models being increasingly used, there has been an interest in developing interpretable models. However, there have been relatively few experimental studies investigating whether these models achieve their intended effects. Thus, the development of new models to provide better interpretability in GBA environments and their validation is still an open challenge.

We found several studies that described data sample and validation challenges. Since most evaluations are conducted with small samples, typically corresponding to one classroom's size, these studies present low statistical power, having a reduced chance of detecting actual effects [123]. Thus, studies must use larger data samples to improve the results' generalization and validity. However, collecting large samples of in-context data is also a cumbersome task. Finally, a few empirical studies discussed the challenge of implementing GBA in the classroom, but this is a significant problem. Many teachers are still unsure about how to integrate game activities with the regular curriculum, and it is crucial to provide guidelines that can facilitate teachers to deploy games in the classroom more easily and flexibly [124].

C. Limitations and implications

This review is mainly limited by the paper selection. First of all, we have only used the key term "game-based assessment" to perform our document search, based on the papers' keywords and titles. However, other communities could also be working on games for assessment purposes, but they might be using slightly different terms to describe their work. Therefore, those studies might not be included in our review. Nevertheless, we purposely opted for this term to analyze the core of GBA while also having a manageable selection of papers for this study. Furthermore, we focused our attention on Scopus and the Web of Science, the two primary academic databases. However, there could be other peer-reviewed academic papers indexed in different databases, as well as non-peer-reviewed publications including pre-prints, technical or white reports that could be missing in our review, and also non-academic work being conducted in industrial companies and by practitioners. Regarding the computational methods and algorithms used, we have identified a set of categories based on the qualitative review of each selected paper. However, there might be studies using less quantitative approaches that might be missing in this review due to the review methodology itself. Finally, we have based our RQ generation on a simplified process that involves the general steps required in GBA projects, but there might be other potential and valuable RQs about the GBA field missing in this review.

We found that most studies emphasized GBA implementation and comparisons between games and classic assessment methods. More studies are needed to systematically develop and improve game design, adopting design-based research methods, as mentioned in [125]. The potential of GBA is now emerging, coinciding with the rise of big data. Data mining and visualization techniques on player interaction logs can provide different stakeholders with valuable insights into how players interact with the game [126]. The increasing interest in games as a learning tool also indicates their potential as actual assessment tools. In our review, we found that GBAs are not only being used in K-16 education but also in medical and professional areas, among others. As expected, the most frequent area where GBAs are being applied is `K-16 education` since children and adolescents are the leading groups whose development will be affected by technology.

Despite this dominating use in education, we can see the great potential that GBAs have in many other contexts. Concerning the professional environment, companies have begun to include assessment games for the recruitment of staff and the selection process. This is a relatively new trend due to certain limitations, such as the cross-domain generalizability of behaviors between game and workforce contexts, which needs further research [40]. In medical environments, the use of GBAs can also be helpful for multiple purposes. Some examples are the possibility to recreate a virtual environment with daily life activities, allowing a precise and complete cognitive evaluation, which can be useful to treat certain diseases such as Alzheimer's [71] or using games to rehabilitate children with cerebral visual impairment using an eye-tracker [94]. Due to the above, we firmly believe that the future of games for assessment is promising; however, further research is needed to overcome the existing problems, and increase the still limited application of games in real-life environments, in order to start building the classrooms of the future.

V. Conclusions

Technology is changing and improving every day, and this is also making a significant impact on educational areas. Moreover, playing games is one of the most popular activities in the world, and the technological revolution that

we are experiencing allows the implementation of games as alternative assessment tools in educational environments. However, previous studies suggest that the use of games also presents some challenges, such as finding the time for both the presenter/instructor and student to learn the systems employed, the financial impact on both parties, and technical limitations [1], [6]. We can tackle all these challenges by facing current limitations and revealing the great potential games have for assessment. This study represents a novel analysis and the first literature review of the emerging research field of GBA. Its main purpose was to review empirical studies of digital GBAs published until 2020. Based on a detailed systematic review of the 65 selected papers, we concluded that games are mainly used in K-16 education for assessment and validation purposes. The domain of the games used is usually related to STEM and cognitive skills, but other domains emerged from our analysis, such as social sciences and physiological capacities. Moreover, we note that, although few GBA studies had the purpose of proposing an adequate game design for assessment, most studies used games designed specifically for assessment purposes, employing complex game-design elements such as collaboration, narrative, or role-playing. In addition, we found that most of the studies used small data samples and simple techniques to process these data and assess students. Finally, we found that most of the studies do not provide public access to their tools, or they overlook links and let them expire over time, which makes it impossible to reproduce the results or even try their game.

Future work should address the current challenges emerging from our review, as those are the main barriers to actual systematic adoption of games for assessment. For example, the next generation of GBA studies should ensure that enough data is collected to have meaningful and reliable results since one of the main limitations of the current research was the size of the data sample collected. Moreover, they should also address the game design that will be used for assessment, as many studies use games designed for other purposes (e.g., entertainment) and overlook the vital link between the design of a game and collecting the necessary evidence for the assessment. In that sense, it would be good to work on conceptual GBA pieces or frameworks that can provide a set of guidelines for the design. Moreover, classic performance indicators such as completion times or scores could still be included in future studies, but GBA also needs to apply more specific and complex algorithms (e.g. knowledge inference or deep learning techniques) specifically designed for learner modeling and assessment purposes. The use of more complex techniques, along with larger data samples, could substantially improve the reliability and generalization of the results. We also believe that future studies should continue exploring the use of visualizations and dashboards to integrate games in schools, adopting a more intuitive approach rather than providing teachers with raw numerical outputs or metrics, which are usually harder to understand. Teachers should also have a more important role in future work to address digital and assessment literacy issues, as well as the potential interpretability and actionability of GBAs. Finally, there are no theoretical frameworks within the GBA area (a related one regarding serious games could be [127]). Considering this lack of theoretical papers focused on describing GBA foundations, we believe that future work should address publications with additional content on the theoretical side.

Therefore, further research is needed to overcome current limitations and to continue exploring the possibilities of games as assessment tools in other contexts and environments. Finally, we encourage authors to document their research in a reproducible and verifiable way, using beneficial open science practices by pre-registering their study, sharing data and code for replication purposes, and if possible open sourcing the GBA tools with clear descriptions so that they can be used by interested stakeholders and researchers.

REFERENCES

[1] L. S. Eiland and T. J. Todd, "Considerations when incorporating technology into classroom and experiential teaching," *The Journal of Pediatric Pharmacology and Therapeutics*, vol. 24, no. 4, pp. 270–275, 2019.

[2] S. De Freitas, "Are games effective learning tools? a review of educational games," *Journal of Educational Technology & Society*, vol. 21, no. 2, pp. 74–84, 2018.

[3] ESA, "2020 essential facts about the computer and video game industry," Entertainment Software Association, Tech. Rep., 2020.

[4] ISFE, "Isfe key facts 2020," ISFE, Tech. Rep., 2020.

[5] S. Papadakis, "The use of computer games in classroom environment," *International Journal of Teaching and Case Studies*, vol. 9, no. 1, pp. 1–25, 2018.

[6] S. de Klerk and P. M. Kato, "The future value of serious games for assessment: Where do we go now?" *Journal of Applied Testing Technology*, vol. 18, no. S1, pp. 32–37, 2017.

[7] D. Clow, "An overview of learning analytics," *Teaching in Higher Education*, vol. 18, no. 6, pp. 683–695, 2013.

[8] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[9] Y. J. Kim and V. J. Shute, "Opportunities and challenges in assessing and supporting creativity in video games," in *Video games and creativity*. Elsevier, 2015, pp. 99–117.

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

[11] K. Dicerbo, "Game-based assessment of persistence," *Educational Technology and Society*, vol. 17, no. 1, pp. 17–28, 2013.

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

[13] D. Ifenthaler, D. Eseryel, and X. Ge, "Assessment for game-based learning," in *Assessment in game-based learning*. Springer, 2012, pp. 1–8.

[14] R. Mislevy, S. Corrigan, A. Oranje, K. DiCerbo, M. Bauer, A. Von Davier, and M. John, *Psychometrics and game-based assessment*, 2015.

[15] Y. J. Kim and D. Ifenthaler, "Game-based assessment: The past ten years and moving forward," in *Game-Based Assessment Revisited*. Springer, 2019, pp. 3–11.

[16] M. Qian and K. R. Clark, "Game-based learning and 21st century skills: A review of recent research," *Computers in human behavior*, vol. 63, pp. 50–58, 2016.

[17] C. Alonso-Fernandez, A. Calvo-Morata, M. Freire, I. Martinez-Ortiz, and B. Fernández-Manjón, "Applications of data science to game learning analytics data: A systematic literature review," *Computers & Education*, vol. 141, p. 103612, 2019.

[18] G. Gris and C. Bengtson, "Assessment measures in game-based learning research: a systematic review," *International Journal of Serious Games*, vol. 8, no. 1, pp. 3–26, 2021.

[19] X. Guan, C. Sun, G.-j. Hwang, K. Xue, and Z. Wang, "Applying game-based learning in primary education: a systematic review of journal publications from 2010 to 2020," *Interactive Learning Environments*, pp. 1–23, 2022.

[20] P.-Y. Chen, G.-J. Hwang, S.-Y. Yeh, Y.-T. Chen, T.-W. Chen, and C.-H. Chien, "Three decades of game-based learning in science and mathematics education: an integrated bibliometric analysis and systematic review," *Journal of Computers in Education*, pp. 1–22, 2021.

[21] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and J. E. McKenzie, "Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n160

[22] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *Bmj*, vol. 372, 2021.

[23] K. S. Tekinbas and E. Zimmerman, *Rules of play: Game design fundamentals*. MIT press, 2003.

[24] B. Kang, S. H. Tan *et al.*, "Interactive games: Intrinsic and extrinsic motivation, achievement, and satisfaction," *Journal of Management and Strategy*, vol. 5, no. 4, pp. 110–116, 2014.

[25] S. De Freitas, *Learning in immersive worlds: A review of game-based learning*. Jisc, 2006.

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

[27] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining" gamification"," in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 2011, pp. 9–15.

[28] H. Al Fatta, Z. Maksom, and M. H. Zakaria, "Game-based learning and gamification: Searching for definitions," *International Journal of Simulation: Systems, Science and Technology*, vol. 19, no. 6, pp. 41–1, 2018.

[29] I. Ghergulescu and C. H. Muntean, "Measurement and analysis of learner's motivation in game-based e-learning," in *Assessment in game-based learning*. Springer, 2012, pp. 355–378.

[30] J. A. Ruipérez-Valiente, "Unveiling the potential of learning analytics in game-based learning: Case studies with a geometry game," in *Handbook of Research on Promoting Economic and Social Development Through Serious Games*. IGI Global, 2022, pp. 524–544.

[31] A. Aghaei Chadegani, H. Salehi, M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, and N. Ale Ebrahim, "A comparison between two main academic literature collections: Web of science and scopus databases," *Asian social science*, vol. 9, no. 5, pp. 18–26, 2013.

[32] Elsevier, "About scopus," 2021. [Online]. Available: https://www.elsevier.com/es-es/solutions/scopus

[33] Clarivate, "Web of science," 2021. [Online]. Available: https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[34] A. Medelyan, "Coding qualitative data: How to code qualitative research," 2021. [Online]. Available: https://getthematic.com/insights/coding-qualitative-data/

[35] M. J. Gomez, J. Ruipérez-Valiente, and F. J. García Clemente, "Supplementary materials: A systematic literature review of game-based assessment empirical studies: Current trends and open challenges," December 2021, https://osf.io/34jk9/?view_only=865f94046fa84c45a013d986bb1c4f87, Accessed November 28, 2022.

[36] K. Di Cerbo, M. Bertling, S. Stephenson, Y. Jia, R. Mislevy, M. Bauer, and G. Jackson, *An application of exploratory data analysis in the development of game-based assessments*, 2015.

[37] J.-P. Lindenmayer, A. Goldring, S. Borne, A. Khan, R. Keefe, B. Insel, A. Thanju, I. Ljuri, and B. Foreman, "Assessing instrumental activities of daily living (iadl) with a game-based assessment for individuals with schizophrenia," *Schizophrenia Research*, vol. 223, pp. 166–172, 2020.

[38] S. Wiloth, N. Lemke, C. Werner, and K. Hauer, "Validation of a computerized, game-based assessment strategy to measure training effects on motor-cognitive functions in people with dementia," *JMIR Serious Games*, vol. 4, no. 2, p. e5696, 2016.

[39] A. B. Collmus, M. B. Armstrong, and R. N. Landers, "Game-thinking within social media to recruit and select job candidates," in *Social media in employee selection and recruitment*. Springer, 2016, pp. 103–124.

[40] E. Short and N. Weidner, "Gamers at work: Predicting workplace-relevant behaviours across domains," *Journal of Gaming and Virtual Worlds*, vol. 11, no. 2, pp. 161–177, 2019.

[41] C. G. I. A. Stanciu and A. T. D. F. Stănescu, "Development of an integrated game based assessment approach–the next generation of psychometric testing," *European Journal of Sustainable Development*, vol. 8, no. 5, pp. 270–270, 2019.

[42] H. Hummel, D. Joosten-ten Brinke, R. Nadolski, and L. Baartman, "Content validity of game-based assessment: case study of a serious game for ict managers in training," *Technology, Pedagogy and Education*, vol. 26, no. 2, pp. 225–240, 2017.

[43] A. Marengo and A. Pagano, "Innovative ways to assess soft-skills: the in-basket game online experience," in *European Conference on e-Learning*. Academic Conferences International Limited, 2020, pp. 325–XVII.

[44] M. Ventura and V. Shute, "The validity of a game-based assessment of persistence," *Computers in Human Behavior*, vol. 29, no. 6, pp. 2568–2572, 2013.

[45] E. Weiner and D. Sanchez, "Cognitive ability in virtual reality: Validity evidence for vr game-based assessments," *International Journal of Selection and Assessment*, vol. 28, no. 3, pp. 215–235, 2020.

[46] M. Cutumisu, K.-L. Turgeon, T. Saiyera, S. Chuong, L. González Esparza, R. MacDonald, and V. Kokhan, "Eye tracking the feedback assigned to undergraduate students in a digital assessment game," *Frontiers in Psychology*, vol. 10, 2019.

[47] R. Almond, "Tips and tricks for building bayesian networks for scoring game-based assessments," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9505, pp. 250–263, 2015.

[48] L. Rivera and C. Suescún, "Game-based assessment for radiofrequency circuits courses in engineering," in *Proceedings - Frontiers in Education Conference, FIE 2015*, vol. 2014, 2015.

[49] F.-Y. Chiu and M.-L. Hsieh, "Role-playing game based assessment to fractional concept in second grade mathematics," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 4, pp. 1075–1083, 2017.

[50] S. Basu, B. Disalvo, D. Rutstein, Y. Xu, J. Roschelle, and N. Holbert, "The role of evidence centered design and participatory design in a playful assessment for computational thinking about data," in *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE 2020*, 2020, pp. 985–991.

[51] Y. Song and J. Sparks, "Measuring argumentation skills through a game-enhanced scenario-based assessment," *Journal of Educational Computing Research*, vol. 56, no. 8, pp. 1324–1344, 2019.

[52] Learningrx, "What are cognitive skills?" 2021. [Online]. Available: https://www.learningrx.com/what-is-brain-training-/what-are-cognitive-skills-/

[53] I. E. Team, "Cognitive skills: What they are and how to improve them," 2021. [Online]. Available: https://www.indeed.com/career-advice/career-development/cognitive-skills-how-to-improve-them

[54] J. Dixon, C. Belnap, C. Albrecht, and K. Lee, "The importance of soft skills," *Corporate finance review*, vol. 14, no. 6, p. 35, 2010.

[55] Y. Song and J. Sparks, "Building a game-enhanced formative assessment to gather evidence about middle school students' argumentation skills," *Educational Technology Research and Development*, vol. 67, no. 5, pp. 1175–1196, 2019.

[56] I. Nikolaou, K. Georgiou, and V. Kotsasarlidou, "Exploring the relationship of a gamified assessment with performance," *Spanish Journal of Psychology*, 2019.

[57] E. M. Mosalam, G. A. El Khayat, S. Lazem, L. Cheniti-Belcadhi, and B. Said, "Assessing modelling readiness in a games environment," in *2019 7th International conference on ICT & Accessibility (ICTA)*. IEEE, 2019, pp. 1–6.

[58] C. Rodríguez-de Pablo, A. Savić, and T. Keller, "Game-based assessment in upper-limb post-stroke telerehabilitation," *Biosystems and Biorobotics*, vol. 15, pp. 413–417, 2017.

[59] V. Korhonen, H. Räty, and E. Kärnä, "A pilot study: a computer game-based assessment of visual perspective taking of four children with autism with high support needs," *Scandinavian Journal of Disability Research*, vol. 19, no. 4, pp. 281–294, 2017.

[60] G. Tanner Jackson, B. Lehman, and L. Grace, "Awkward annie: Impacts of playing on the edge of social norms," in *ACM International Conference Proceeding Series 2020*, 2020.

[61] G. Jackson, L. Grace, P. Inglese, J. Wain, and R. Hone, "Awkward annie: Game-based assessment of english pragmatic skills," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10714 LNCS, pp. 795–808, 2018.

[62] H. Song, D.-J. Yi, and H.-J. Park, "Validation of a mobile game-based assessment of cognitive control among children and adolescents," *PLoS ONE*, vol. 15, no. 3, 2020.

[63] H. Ketamo and K. Devlin, "Replacing pisa with global game based assessment," in *Proceedings of the European Conference on Games-based Learning 2014*, vol. 1, 2014, pp. 258–264.

[64] R. Halverson and V. Owen, "Game-based assessment: An integrated model for capturing evidence of learning in play," *International Journal of Learning Technology*, vol. 9, no. 2, pp. 111–138, 2014.

[65] O. Gaggi, T. Sgaramella, L. Nota, M. Bortoluzzi, and S. Santilli, "A serious games system for the analysis and the development of visual skills in children with cvi: A pilot study with kindergarten children," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 195 LNICST, pp. 155–165, 2017.

[66] M. Bertling, G. Tanner Jackson, A. Oranje, and V. Owen, "Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9112, pp. 545–549, 2015.

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

[68] M. Loachamin-Valencia, M.-C. Juan, M. Mendez-Lopez, and E. Perez-Hernandez, "Auditory and spatial assessment in inattentive children using smart devices and gesture interaction," in *Proceedings - IEEE 17th International Conference on Advanced Learning Technologies, ICALT 2017*, 2017, pp. 106–110.

[69] S. Pouezevara, S. Powers, G. Moore, C. Strigel, and K. McKnight, "Assessing soft skills in youth through digital games," in *ICERI2019 Proceedings*. IATED, 2019, pp. 3057–3066.

[70] A. Mavridis and T. Tsiatsos, "Game-based assessment: investigating the impact on test anxiety and exam performance," *Journal of Computer Assisted Learning*, vol. 33, no. 2, pp. 137–150, 2017.

[71] V. Vallejo, P. Wyss, L. Rampa, A. Mitache, R. Müri, U. Mosimann, and T. Nef, "Evaluation of a novel serious game based assessment tool for patients with alzheimer's disease," *PLoS ONE*, vol. 12, no. 5, 2017.

[72] A. Leonardou, M. Rigou, and J. Garofalakis, "Techniques to motivate learner improvement in game-based assessment," *Information (Switzerland)*, vol. 11, no. 4, 2020.

[73] V. Abeele, J. Wouters, P. Ghesquière, A. Goeleven, and L. Geurts, "Game-based assessment of psychoacoustic thresholds: Not all games are equal!" in *CHI PLAY 2015 - Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 2015, pp. 331–342.

[74] K. Gjicali, B. Finn, and D. Hebert, "Effects of belief generation on social exploration, culturally-appropriate actions, and cross-cultural concept learning in a game-based social simulation," *Computers and Education*, vol. 156, 2020.

[75] J. Hautala, R. Heikkilä, L. Nieminen, V. Rantanen, J.-M. Latvala, and U. Richardson, "Identification of reading difficulties by a digital game-based assessment technology," *Journal of Educational Computing Research*, vol. 58, no. 5, pp. 1003–1028, 2020.

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

[77] J. Perry, S. Balasubramanian, C. Rodriguez-De-Pablo, and T. Keller, "Improving the match between ability and challenge: Toward a framework for automatic level adaptation in game-based assessment and training," in *IEEE International Conference on Rehabilitation Robotics*, 2013.

[78] M. Gómez-Álvarez, J. Echeverri, and L. González-Palacio, "Games-based assessment strategy: Case systems engineer of universidad de medellín [estrategia de evaluación basada en juegos: Caso ingeniería de sistemas universidad de medellín]," *Ingeniare*, vol. 25, no. 4, pp. 633–642, 2017.

[79] Y. Jaffal and D. Wloka, "Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content," *Journal of E-Learning and Knowledge Society*, vol. 11, no. 3, pp. 101–115, 2015.

[80] M. Cutumisu, D. Chin, and D. Schwartz, "A digital game-based assessment of middle-school and college students' choices to seek critical feedback and to revise," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 2977–3003, 2019.

[81] D. Chin, K. Blair, and D. Schwartz, "Got game? a choice-based learning assessment of data literacy and visualization skills," *Technology, Knowledge and Learning*, vol. 21, no. 2, pp. 195–210, 2016.

[82] C. Forsyth, T. Jackson, D. Hebert, B. Lehman, P. Inglese, and L. Grace, "Striking a balance: user-experience and performance in computerized game-based assessment," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10331 LNAI, pp. 502–505, 2017.

[83] C. Arce-Lopera and A. Perea, "Logic evaluation through game-based assessment," *Advances in Intelligent Systems and Computing*, vol. 973, pp. 243–250, 2020.

[84] L. Tari, *Knowledge Inference*. New York, NY: Springer New York, 2013, pp. 1074–1078.

[85] V. Shute and L. Wang, *Assessing and Supporting Hard-to-Measure Constructs in Video Games*, 2016.

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

[87] M. Hargrave, "Deep learning," 2020. [Online]. Available: https://www.investopedia.com/terms/d/deep-learning.asp/

[88] F. Chen, Y. Cui, and M.-W. Chu, "Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 3, pp. 481–503, 2020.

[89] G. Dong and J. Pei, *Sequence data mining*. Springer Science & Business Media, 2007, vol. 33.

[90] M. J. Gomez, J. A. Ruipérez-Valiente, P. A. Martinez, and Y. J. Kim, "Exploring the affordances of sequence mining in educational games," in *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2020, pp. 648–654.

[91] Oxford, "Oxford languages and google," 2021. [Online]. Available: https://languages.oup.com/google-dictionary-en/

[92] M. Lonergan, L. De Wet, and A. Burger, "Technology as a tool to improve student understanding of assessment questions," *Advances in Intelligent Systems and Computing*, vol. 1217 AISC, pp. 250–256, 2020.

[93] F.-H. Tsai, C.-C. Tsai, and K.-Y. Lin, "The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment," *Computers and Education*, vol. 81, pp. 259–269, 2015.

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

[95] J. Delcker and D. Ifenthaler, "Mobile game-based language assessment," *International Journal of Emerging Technologies in Learning*, vol. 15, no. 3, pp. 195–206, 2020.

[96] P. Martínez, M. Gómez, J. Ruipérez-Valiente, G. Pérez, and Y. Kim, "Visualizing educational game data: A case study of visualizations to support teachers," in *CEUR Workshop Proceedings*, vol. 2671, 2020, pp. 97–111.

[97] J. Paiva and J. Leal, "Asura: A game-based assessment environment for mooshak," in *OpenAccess Series in Informatics*, vol. 62, 2018.

[98] C. Harteveld and S. Sutherland, "The goal of scoring: Exploring the role of game performance in educational games," in *Conference on Human Factors in Computing Systems 2015 - Proceedings*, vol. 2015-April, 2015, pp. 2235–2244.

[99] D. Jackson, S. Kim, C. Lee, Y. Choi, and J. Song, "Simulating déjà vu: What happens to game performance when controlling for situational features?" *Computers in Human Behavior*, vol. 55, pp. 796–803, 2016.

[100] B. Lehman, D. Hebert, G. Jackson, and L. Grace, "Affect and experience: Case studies in games and test-taking," in *Conference on Human Factors in Computing Systems 2017 - Proceedings*, vol. Part F127655, 2017, pp. 917–924.

[101] J. Hao and R. Mislevy, "The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design," *ETS Research Report Series*, vol. 2018, no. 1, 2018.

[102] H.-H. Yu, J.-K. Yang, H.-W. Chen, T.-F. Yu, and H.-T. Hou, "Jingnan campaign© - using game-based assessment with the mechanism of strategy games for history teaching: System development and learning evaluation," in *Proceedings - 2015 IIAI 4th International Congress on Advanced Applied Informatics, IIAI-AAI 2015*, 2016, pp. 727–728.

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

[104] E. Ibryamova and G. Stefanov, "Developing and implementing a labyrinth game for self-assessment," in *ACM International Conference Proceeding Series*, 2020, pp. 106–110.

[105] D. Sanchez and M. Langer, "Video game pursuit (vgpu) scale development: Designing and validating a scale with implications for game-based learning and assessment," *Simulation and Gaming*, vol. 51, no. 1, pp. 55–86, 2020.
[106] M. Ponticorvo, F. Ferrara, R. Di Fuccio, A. Di Ferdinando, and O. Miglino, "Sniff: A game-based assessment and training tool for the sense of smell," *Advances in Intelligent Systems and Computing*, vol. 617, pp. 126–133, 2017.
[107] A. S. Robberts and L. Van Ryneveld, "Design principles for introducing 21st century skills by means of game-based learning," *Industry and Higher Education*, p. 09504222221079210, 2022.
[108] M. J. Dondlinger, "Educational video game design: A review of the literature," *Journal of applied educational technology*, vol. 4, no. 1, pp. 21–31, 2007.
[109] J. P. Gee, "Are video games good for learning?" *Nordic Journal of Digital Literacy*, vol. 1, no. 03, pp. 172–183, 2006.
[110] M. Akcaoglu and M. J. Koehler, "Cognitive outcomes from the game-design and learning (gdl) after-school program," *Computers & Education*, vol. 75, pp. 72–81, 2014.
[111] M. H. Hussein, S. H. Ow, M. M. Elaish, and E. O. Jensen, "Digital game-based learning in k-12 mathematics education: a systematic literature review," *Education and Information Technologies*, pp. 1–33, 2021.
[112] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2019.
[113] K. Cukier, "A special report on managing information," *The Economist*, vol. 394, no. 8671, pp. 3–18, 2010.
[114] J. Chin, R. Dukes, and W. Gamson, "Assessment in simulation and gaming: A review of the last 40 years," *Simulation & Gaming*, vol. 40, no. 4, pp. 553–568, 2009.
[115] R. J. Mislevy and G. D. Haertel, "Implications of evidence-centered design for educational testing," *Educational measurement: issues and practice*, vol. 25, no. 4, pp. 6–20, 2006.
[116] V. J. Shute, "Stealth assessment in computer-based games to support learning," *Computer games and instruction*, vol. 55, no. 2, pp. 503–524, 2011.
[117] C. I. Johnson and R. E. Mayer, "Applying the self-explanation principle to multimedia learning in a computer-based game-like environment," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1246–1252, 2010, online Interactivity: Role of Technology in Behavior Change.
[118] A. All, E. P. N. Castellar, and J. Van Looy, "Measuring effectiveness in digital game-based learning: A methodological review." *International Journal of Serious Games*, vol. 1, no. 2, 2014.
[119] F. Fidler and J. Wilcox, "Reproducibility of scientific results," *The Stanford Encyclopedia of Philosophy*, 2018.
[120] T. van der Zee and J. Reich, "Open education science," *AERA Open*, vol. 4, no. 3, p. 2332858418787466, 2018.
[121] S. Buck, "Solving reproducibility," 2015.
[122] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–52.
[123] G. Petri and C. G. von Wangenheim, "How games for computing education are evaluated? a systematic literature review," *Computers & education*, vol. 107, pp. 68–90, 2017.
[124] M. J. Gomez, J. A. Ruipérez-Valiente, P. A. Martínez, and Y. J. Kim, "Applying learning analytics to detect sequences of actions and common errors in a geometry game," *Sensors*, vol. 21, no. 4, p. 1025, 2021.
[125] M.-C. Li and C.-C. Tsai, "Game-based learning in science education: A review of relevant research," *Journal of Science Education and Technology*, vol. 22, no. 6, pp. 877–898, 2013.
[126] M. Freire, Á. Serrano-Laguna, B. Manero, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, "Game learning analytics: learning analytics for serious games," in *Learning, design, and technology*. Springer Nature Switzerland AG, 2016, pp. 1–29.
[127] C. S. Loh, Y. Sheng, and D. Ifenthaler, "Serious games analytics: Theoretical framework," in *Serious games analytics*. Springer, 2015, pp. 3–29.

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

José A. Ruipérez-Valiente (Senior Member, IEEE) received his B.Eng. degree in telecommunications from Universidad Católica de San Antonio de Murcia, and a M.Eng. degree in telecommunications together with his M.Sc. and Ph.D. degrees in telematics from Universidad Carlos III of Madrid while conducting research with Institute IMDEA Networks in the area of learning analytics and educational data mining. He was a postdoctoral associate at MIT. He has received more than 20 academic/research awards and fellowships, has published more than 100 scientific publications in high impact venues, and participated in over 18 funded projects. He is currently an Assistant Professor of Computer Science and Artificial Intelligence at the University of Murcia.

Félix J. García Clemente is Associate Professor at the Department of Computer Engineering, UMU. García Clemente received his M.Sc. and Ph.D. degrees in Computer Science from the University of Murcia, Spain. His teaching includes courses in computer networks, network management, ubiquitous computing, and mobile device programming. His major research interests focus on Cybersecurity, Distributed Management of Networks and Services and Interaction Systems.