A SYSTEMATIC LITERATURE REVIEW OF DIGITAL GAME-BASED ASSESSMENT EMPIRICAL STUDIES: CURRENT TRENDS AND OPEN CHALLENGES

Manuel J. Gomez, José A. Ruipérez-Valiente, Félix J. García Clemente
Faculty of Computer Science
University of Murcia
Murcia (Spain)
{manueljesus.gomezm, jruiperez, fgarcia}@um.es

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, based on 66 research papers that used digital GBAs to determine: (1) the context where the study has been applied, (2) the primary purpose, (3) the knowledge domain of the game used, (4) game/tool availability, (5) the size of the data sample, (6) the data science techniques 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 formal education and for assessment purposes, and most GBAs focus on assessing STEM content and cognitive 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 the status of the field with the current trends and the open challenges (including replication and validation problems) providing recommendations for the future research agenda of the GBA field.

Keywords Game-based assessment · educational technology · game-based learning · learning analytics

1 Introduction
Technology is progressively changing the world in which we live in. 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 in education [2]. Digital games have become a significant part of families and, specially, young people around the world. In fact, three quarters of all U.S. households have at least one person who plays video games [3], meanwhile in Europe, 51% of the population aged 6-61 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 the classic methods of assessment do not seem to fully have the power to measure all aspects of students’ knowledge, skills, and attributes [6].

Accompanying this technology use explosion 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 of e-learning resources, instrumental software, and the use of the Internet have created large repositories that provide a goldmine of educational data that can be explored and exploited to understand how students learn [8]. Regarding games, they allow to recreate more realistic
situations compared to classic classroom activities. From these situations, we can collect a vast amount of detailed data about student’s 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 competences acquired, most traditional tests present individual and disconnected items to learners, while 21st-century competencies benefit from being applied within the context for more accurate measurements. Furthermore, classic assessment often interrupts the learning process, and it does little to motivate learners [11]. Because 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 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 [13]. So, how is the current state of the art in the GBA area for effective implementations? 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, authors in [14] reviewed 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, All et al. [15] conducted a survey focusing on measuring effectiveness in digital GBL, and, similarly, Li & Tsai [16] reviewed 31 articles from 2000 to 2011 to analyze relevant research of GBL in science education. Finally, we found that Alonso-Fernandez et al. [17] performed 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. However, we have not found any specific study reviewing literature about GBA. The current paper aims to conduct the first systematic literature review on the applications of empirical GBA studies and answer a set of research questions based on the analysis performed to discover current trends and open challenges in this area. These results will provide an overall view of the GBA field, defining its current status and potential future steps in the research of this area.

The rest of the paper is organized as follows. Section 2 describes the methods, including the research questions, databases and search terms, research selection, and review process. Section 3 presents the results synthesis and analysis. Then, we finalize the paper with discussion in Section 4 and conclusions in Section 5.

2 Methods

We followed a standard systematic literature review methodology, using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [18] as a basis for reporting our systematic review. First of all, (1) we stated each research question (RQ), (2) we used a fixed set of queries on a pre-identified bibliographical database, and (3) a set of inclusion and exclusion criteria. Then, (4) we made a full paper review and coding process of the RQs, and (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 [19]) in Figure 1.

2.1 Research questions

To state each one of the research questions, we analyzed and simplified the steps in a GBA empirical research [20], which can be seen in Figure 2. From these steps, we identified the following research questions, which will allow us to understand the open challenges and current trends in the area:

RQ1. In which context or environment has the GBA been applied?

RQ2. Which is the primary purpose of the GBA?

RQ3. Which is the knowledge domain of the GBA?

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

RQ5. Which is the size of the data sample used in the study?

RQ6. What data science techniques 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?
A Systematic Literature Review of Digital GBA Empirical Studies

2.2 Databases and search terms

We have queried two databases: Scopus and Web of Science, since they are the most widespread databases on different scientific fields, being frequently used for surveying the literature [21]. 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 [22]. Moreover, Web of Science is the second biggest bibliographic database, allowing to track ideas across disciplines and time from almost 1.9 billion cited references from over 171 million records [23].

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

Figure 2: A simplified view of the steps in game-based assessment empirical research.
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 following final search query:

\[(\text{TITLE} ("game-based assessment") \text{ OR KEY} ("game-based assessment"))\].

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

2.3 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. After this first analysis, we classified studies as included or excluded. The inclusion/exclusion criteria followed is described below. Accordingly, 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 authors have high proficiency): 0% of the papers were excluded.
- We were able to access the full text of the paper: 1.9% of the papers were excluded (2 studies).
- The paper was published in a conference, journal or as a 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: 4.2% of the papers were excluded (4 studies).
- The paper included empirical evidence related to the outcomes of applying the GBA tool: 27.5% of the papers were excluded (25 studies).

2.4 Final paper collection

After the first brief review to ensure that every paper met our inclusion/exclusion criteria, we excluded a total of 38 papers. Thus, the final paper collection consists of 66 studies.

Figure 3 shows the distribution of papers within the final collection by their publication year. We clearly see an increasing interest on this particular topic: between years 2013 and 2016, we only have 21 published papers that matched our criteria, while between 2017 and 2020 we have 45 of them.

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 obviously the most common one. The most frequent keywords are presented in Figure 4. The total sum of keywords is 304 while there are 203 unique keywords. The average keyword was found 1.5 times. As we can see the predominant keywords are strongly focused on games, assessment, and analytics.

2.5 Review and coding process

In the coding stage, we collected the data of the selected studies that we consider as the most valuable to address the RQs defined in Section 2.1. Based on the aim of the review, we followed an inductive coding scheme (also called open coding), meaning that codes created were based on the qualitative data itself [24]. This is an iterative process, since you can add new codes, split an existing code into two, or compress two existing codes into one as you continue reviewing data. Specifically, in our analysis we first made a brief review of each paper, collecting all the necessary information to code each RQ at once. After that, we made an iterative process where we continued reviewing the information corresponding to each RQ sequentially. The whole results of the coding process per paper are available at [25]. In addition, note that each paper can fit into more than one of the codes created for each RQ.

3 Results

3.1 In which 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. **Formal education**: these papers use GBAs in formal education (e.g., school, university) to support teaching and learning. More specifically, games are most commonly used in middle school and high school (22.7%). However, games are also used in other formal education environments such as primary school (15.2%) and university (12.1%). For example, Di Cerbo et al. [26] used game data from 751 US middle school players, and Lin et al. [27] evaluated two third-year university courses in Taiwan.

2. **Medical**: games can also be used in medical environments with different purposes (e.g., rehabilitation). For example, authors in [28] examined the feasibility of administering the GBA in a sample of inpatients with chronic schizophrenia with low levels of functioning, meanwhile authors in [29] 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, authors tested eight hypotheses, being seven of them confirmed (87.5%), thus indicating an excellent construct validity. Moreover, authors found moderate-to-high test–retest reliability (ICC=0.47-0.92).

3. **Workplace**: another option is the use of games to assess in professional environments. This way, enterprises can use these games to evaluate their employees or provide them with additional feedback. Even now, companies can include the use of GBA for the employee recruitment and selection process [30]. This idea is supported by the fact that in-game constructs show similar relationships with in-game performance as the workplace constructs do with job performance [31].

There are also some studies, such as [32], that did not specify in which context their games were being used (11 studies, 16.7%). We can see the number of papers fitting in each category in Figure 3. As we can observe, GBAs are mostly used in formal education (42 studies, 63.6%), followed by medical (8 studies, 12.1%) and workplace contexts (5 studies, 7.6%).

### 3.2 Which is the primary purpose of the GBA? (RQ2)

We coded 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 them.

1. **GBA evaluation**: in these studies, authors evaluate the game by checking if it achieves their initial objectives using some measure to prove that the game or tool is suitable to be used in an educational environment. In [33],

![Figure 3: Number of selected studies per year of publication.](image-url)
authors presented 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 face-to-face assessment. Moreover, authors in [34] 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 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 [35] 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 to improve the learning process using this evaluation measure instead of classic evaluation methods or to provide personalized feedback. In their work, Weiner & Sanchez [36] 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 [37], 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, authors propose the design of a new framework to be used within the context of GBA. We can see an example in [38], 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 with assessment purposes. For example, authors in [39] 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 [36] used a virtual reality game to calculate a score measure for each student (assessment) and, moreover, they proved that these calculated scores are appropriate to be used by comparing them to classic measures (GBA evaluation).

We can see the number of papers fitting each category in Figure 5. GBA evaluation is the most common category (39 studies, 59.1%), followed by assessment (35 studies, 53.0%) and framework proposal (12 studies, 18.2%). The less common category is game design proposal, with only three papers fitting (4.5%). 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.

3.3 Which is the knowledge domain of the GBA? (RQ3)

From reviewing the selected papers, we identify four major knowledge domain categories: science, technology, engineering and mathematics (STEM), cognitive skills, humanities and social sciences, and physiological capacities. As some of the categories have more than one related area, we also consider some sub-categories fitting into them. Let us describe each domain category in detail:
1. STEM: in this category, we include papers that are related with science, technology, engineering and mathematics. For example, Chiu & Hsieh [40] showed the different teaching methods of second grade elementary students in fraction concepts (mathematics), while Kim et al. [41] 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 with humanities and social science areas (e.g., art, music, language) fit 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 [42], where authors used a game where 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 with language) in [43], where they described the design of an argumentative reasoning task within a scenario-based assessment enhanced with game elements.

3. Cognitive skills: cognitive skills are the core skills your brain uses to think, read, learn, remember, reason, and pay attention. 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 [44, 45]. In this category we consider attention, memory, logic and reasoning, and visual processing and speed. We find papers that have measured interesting skills, such as [46], which included a series of reasoning activities to measure argumentation skills (which is related with logic and reasoning), or [47, 48], 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. [49], 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 to motivate children with autism to make more eye-contact [50].

Figure 6: Main purpose of each research. More than one purpose is possible for each paper.
There are also papers fitting more than one category at once. For example, in [51, 52], 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 skills. We can see the full tree showing the distribution on studies into categories and sub-categories in Figure 7.

![Category tree for RQ3](image)

Figure 7: Category tree for RQ3.

The three predominant categories are humanities and social sciences (18 studies, 27.3%), cognitive skills (25 studies, 37.9%) and STEM (19 studies, 28.8%). Taking a look at each sub-category, we note that the main area in STEM is science (10 studies, 15.2%). The main field in cognitive skills is logic and reasoning, with 17 fitting papers (25.8%), while in humanities and social sciences, the predominant sub-category is language, with nine papers fitting (13.6%).

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

A critical aspect of research is the availability of the generated results to be used by the public. It is essential to make tools accessible so that researchers can replicate experiments and practitioners can use it as part of their teaching. From our paper analysis, we find three primary categories: Currently available, Not available 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 its public use (e.g., [53, 54]).

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., [55, 56, 57]).
3. **Not specified**: researchers did not specify the tool’s availability (it is more likely that it is not accessible) (e.g., [58, 59, 60]).

![Figure 8: Number of papers based on their availability.](image)

Figure 8 shows that most papers (51 studies, 77.2%) did not specify if the tool is accessible or not. Another minority of papers (10 studies, 15.2%) offered their tool publicly. The rest of the studies (five papers, 7.6%) initially offered their games, but they are currently unavailable.

### 3.5 Which 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. From the coding process, we present four categories:

1. **Less than 50 participants**: these papers involved less than 50 participants in their empirical studies. We find studies with small data samples, such as [61], using data from 30 postgraduate students, or [62], 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. [63] used data from 77 primary school pupils for assessing and improving multiplication skills. We see another example in [64], 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. [65] 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. [66] 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 for reading fluency and 0.87 for word reading, pseudoword reading, word spelling, and pseudoword spelling tasks). 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.

Other papers (e.g., [67, 68]) did not specify the data sample size of the study and we categorize them as **NA**. Figure 9 summarizes the results on the data sizes across papers. We can see that only 16 papers (24.2%) used more than 250
participants in their students, and only nine papers (13.6%) used data from more than 500 students, meanwhile most of the papers (59.1%) used data from less than 250 participants. We also see that a significant amount of papers (11 studies, 16.7%) that did not specify the data sample size in their studies.

3.6 What data science techniques and algorithms have been applied in the research? (RQ6)

After exploring the data sample that were retrieved across papers, our goal was to examine the methods that were applied for its analysis. Accordingly, we identify five different groups of methods for analyzing the data: Descriptive statistics, Machine learning, Knowledge inference, Deep learning, and Data mining. Let us see each group in detail:

1. **Descriptive statistics**: these englobe further mathematical analyses covering a lot of methods, tests, and visualizations. We identified several works that applied summary statistics (e.g., mean, variances) [69], correlations [70] and visualizations [71].

2. **Machine learning**: it is considered as a part of artificial intelligence, and it covers a set of methods that allow systems to learn and improve from historical data automatically. We noted that authors used two significant families of machine learning methods: supervised learning and unsupervised learning. Supervised learning includes techniques such as regression [72], meanwhile unsupervised learning uses other methods, such clustering techniques like $k$-means [12] or dimensionality reduction techniques like Principal Component Analysis (PCA) [73]. For example, authors in [74] 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 conducting a linear regression (which is a supervised method), concluding that the measures obtained from both methods were not significantly different.

3. **Knowledge inference**: it refers to acquiring 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 [75]. Common knowledge inference methods that several studies have used are bayesian networks [76] and fuzzy cognitive maps [58]. In [77], researchers proposed a dynamic Bayesian network modeling approach for measuring student performance from an educational video game. The results supported that Bayesian networks are a useful way for characterizing and accumulating evidence regarding students in games and related assessment environments.

4. **Deep learning**: it is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for decision making [78]. An example is the work of Chen et al. [79], who used Long Short-Term Memory (LSTM), an artificial recurrent neural network architecture.
5. **Data mining**: the objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data [80]. Specifically, Gomez et al. used [81] sequence mining to identify sequences and errors by transforming raw data into meaningful sequences that are interpretable and actionable for teachers.

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 of the papers (83.3%) used *Descriptive statistics*, and almost none of them used *deep learning* (only 1 paper, 1.5%). We also noted that 80% of the papers that used *machine learning* techniques used *supervised learning* too, and specifically most of them used different types of regressions.

### 3.7 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” [82]. In our study, we consider the paper’s stakeholder as the person to whom the results are directed. Specifically, we have two main groups of stakeholders: researchers and 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. [83] created a Paper-based Assessment (PBA) and a GBA in order to measure students’ performance, cognitive states and satisfaction related to both assessment methods, concluding that smaller versatile GBAs may have a greater impact on the student’s cognitive capabilities, and could possibly enhance student performances during, for example, a main exam or short formative assessments. Moreover, Tsai et al. [84] proposed an online learning system using different gaming modes of classic...
tic-tac-toe game to explore how different gaming modes and feedback types in this game-based formative assessment affect knowledge acquisition effectiveness and participation perceptions.

2. **Final user**: if the paper’s results are designed 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, students). In their work, Ciman et al. [85] 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 [86] also developed a mobile app that makes an automated analysis of the data and provides information about children’s language skills. Other works are focused on teachers, such as [87], where authors used a GBA to develop a set of visualizations to support teachers in classrooms.

![Figure 11: Paper distribution tree based on the main stakeholder.](image)

Figure 11 shows the number of studies that focused their work on each of the stakeholders. We see that 34 studies (51.5%) were directed to final users, mainly students. Meanwhile, 30 studies (45.5%) are focused on researchers as the main result recipients. Another two studies (e.g., [88]) 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 (34.8%) and teachers (13.6%).

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

Limitations show potential weak points of the study that researchers usually highlight regarding their own work, due to constraints in research design or methodology, for example. 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**: the correct design of a game is crucial for students’ 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 to the final purpose of the assessment. Authors in [89] informed that many design decisions, in the design of the game and in the design of the training, play a role in what kind of game performance is achieved and its meaning, and future designers should keep this in mind to create cleaner results to achieve more robust measures.

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. [90] reported that they had a small sample size, and that larger sample sizes would be necessary to detect smaller effects in the future. We see a similar example in [91], where authors had a sample collection of 67 students, but only four of those 67 students samples were used in their empirical study. The work in [92] reported 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 schedule.

3. **Methodological**: in this category, we include challenges and limitations related to methods, algorithms, or techniques used. For example, Yu et al. [93] wanted to collect additional data to explore learners’ behavioral patterns during gameplay. We see another example in [94], since 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 the way a machine or system works. That could include storage limitations, computing power, or even limitations related to sensors used in the study. In [95], authors reported the necessity of a database (to store information about students’ achievements), since they could not store that information, and an administrator 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 limitations as the schools’ budget or the rigidity of subjects’ classic curriculum. Halverson & Owen [55] supported 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 were proved satisfactory by conducting empirical experiments. Sanchez & Langer [96] informed 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.

Some other studies did not report any challenge or limitation (e.g., [97]). Figure 12 shows that methodological challenges are the most common ones (34 studies, 51.5%), followed by data sample limitations (31 studies, 47.0%).
Table 1: Summary of the main findings.

| Research question          | Categories                  | Count | %   |
|----------------------------|-----------------------------|-------|-----|
| Context (RQ1)              | Formal education            | 42    | 63.6%|
|                            | Medical                     | 8     | 12.2%|
|                            | Workplace                   | 5     | 7.6% |
|                            | Not Available               | 11    | 16.7%|
| Main Purpose (RQ2)         | GBA evaluation              | 39    | 59.1%|
|                            | In-game behaviors           | 8     | 12.1%|
|                            | Assessment                  | 35    | 53.0%|
|                            | Interventions               | 7     | 10.6%|
|                            | Framework proposal          | 12    | 18.2%|
|                            | Game design proposal        | 3     | 4.5% |
| Knowledge Domain (RQ3)     | STEM                        | 19    | 28.8%|
|                            | Cognitive skills            | 25    | 37.9%|
|                            | Humanities and social sciences | 18    | 27.3%|
|                            | Physiological capacities    | 8     | 12.1%|
|                            | Not Available               | 3     | 4.5% |
| Availability (RQ4)         | Not specified               | 51    | 77.3%|
|                            | Currently available         | 10    | 15.2%|
|                            | Not available               | 5     | 7.6% |
| Sample Size (RQ5)          | Less than 50 participants   | 19    | 28.8%|
|                            | Between 50 and 250 participants | 20    | 30.3%|
|                            | Between 250 and 500 participants | 7     | 10.6%|
|                            | More than 500 participants  | 9     | 13.6%|
|                            | Not available               | 11    | 16.7%|
| Algorithms/techniques (RQ6)| Descriptive statistics      | 55    | 83.3%|
|                            | Machine learning            | 25    | 37.9%|
|                            | Deep learning               | 1     | 1.5% |
|                            | Data mining                 | 1     | 1.5% |
|                            | Knowledge inference         | 7     | 10.6%|
|                            | Not available               | 8     | 12.1%|
| Stakeholder (RQ7)          | Researchers                 | 30    | 45.5%|
|                            | Final user                  | 34    | 51.5%|
|                            | Not available               | 2     | 3.0% |
| Limitations (RQ8)          | Technical                   | 9     | 13.6%|
|                            | Game design                 | 20    | 30.3%|
|                            | Data sample                 | 31    | 47.0%|
|                            | Methodological              | 34    | 51.5%|
|                            | Integration                 | 6     | 9.1% |
|                            | Validation                  | 25    | 37.9%|
|                            | Not available               | 5     | 7.6% |

In the other hand, the rarest limitations are related with integration (6 studies, 9.1%) and technical challenges (9 studies, 13.6%).

4 Discussion

In this section, we first present a summary and discussion of our primary findings. We can also see a summary of the main findings in Table 1. Next, we will present a discussion about past and future challenges regarding games for assessment. Finally, we will talk about existing limitations in this study and the implications of our research.

4.1 Current trends

First of all, we analyzed the contexts where the studies were applied (RQ1), finding that most of them took place in formal education, especially in high school and middle school. This is an interesting finding because young kids and teenagers constitute the major force whose 21st-century competencies development will be heavily impacted...
by technology [14]. Moreover, children and adolescents are an ideal target, since the familiarity of these users with gaming environments and game mechanics facilitate their interactions with games [17].

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. 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 cannot forget about game design. Literature reveals that game design is essential, and several distinct design elements, such as narrative context, rules, goals, rewards, multi-sensory cues, and interactivity, seem necessary to stimulate desired outcomes [98, 99].

We also extracted four predominant knowledge domains (RQ3) across studies. A large proportion of the analyzed studies aimed at practicing and assessing content related to STEM and 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 fraction of papers also focused on developing and measuring cognitive skills. Using game design as a context to teach higher-order thinking skills has received interest from researchers, since schools usually place a heavy emphasis on covering and delivering content knowledge [100]. 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) [62, 59].

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. Concerning the data sample, we also tried to collect information about the type of data collected, but few studies provided that information.

Across papers, researchers used many different algorithms and techniques (RQ6) to analyze the data. We classified them into five categories, finding that the most common ones are descriptive statistics and machine learning. With that said, we note that the majority of papers used simple statistics and machine learning algorithms, such as linear regressions, and few studies used more complex or advanced techniques. Simpler techniques could potentially restrict the significance and generalization of the results, as they might not be adequate to model student’s knowledge properly. However, since those techniques require less technical and statistical expertise, they are easier to apply and implement, which could be the reason why researchers choose them more frequently. Therefore, more work is needed to develop specialized GBA methods, that are also affordable to implement, perhaps through open sourced 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 [101] and because it is easier for the brain to comprehend an image versus words or numbers [102]. We think this is a promising way to integrate games in schools, and we realized that studies are now tending to use visualizations to communicate their results (e.g., [55, 71]).

Finally, we want to report the scarce information regarding games and tools availability (RQ4). The majority of studies did not provide any information of how to access or use their tools. In addition, some studies made tools public but expired, being inaccessible at present. This highlights 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 research.

4.2 Open challenges

From our results and previous related reviews, we find some open challenges in the area that authors usually report. Let us see each one of these challenges in detail:

In [103], authors address the challenge of how to make appropriate assessments. They noted that pre-test and post-test measures are a good way to assess, 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 way to make good assessments using Evidence-centered Design (ECD) and stealth assessment. ECD framework views assessment as an evidentiary argument: an argument from which we observe students say, do, or make in a few particular circumstances, to inferences about what they know or do [104]. Moreover, stealth assessment represents a quiet, yet powerful process by which learner performance data are continuously gathered during the course of playing and learning, and inferences are made about the level of relevant competencies, maintaining the flow and engagement [105]. Since ECD and stealth assessment are two common practices in current research, we could affirm that the objective of making appropriate assessments using unobtrusive methods has been accomplished.
We found that most studies emphasized GBA implementation and comparisons between games and classic assessment. We found that several studies that reported 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. Thus, studies must use larger data samples to improve the results' generalization and validity. However, collecting large samples of in-context data is a cumbersome task as well. Finally, few works discussed about the challenge of implementing GBA in the classroom, but this is actually 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.

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 not available in most cases, All et al. [15], mentioned replication issues with certain studies due to missing information on multiple areas of the study. It is crucial to provide a detailed description of the followed procedure to conduct the study. While the community is currently demanding more standardized open science practices, this problem is currently still present. Besides that, Alonso-Fernandez et al. [17] noted that most papers did not report the format in which they collected the data, so we can not know if they were using a standard or relying on their own data formats, which represent even more replication and reusability issues. 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 the low reproducibility. In fact, the terms “reproducibility crisis” and “replication crisis” have gained a significant popularity over the last decade [107]. To fix this issue, the community is demanding having more pre-registered studies, open data, open analyses, and open access publications [108], and this can be systematized by the guidelines of the publishers, governments and research communities [109].

Regarding methods and techniques, we identify the challenge of implementing learner modeling algorithms. As we mentioned before, studies 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 fact, 50% 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 (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 the low interpretability. Even if we said that visualizations are a promising way to improve the result’s presentation and communication, they can not improve model’s interpretability themselves. According to researchers in [110], 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 a better interpretability in GBA environments and their validation is still an open challenge too.

We found that several studies that reported 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 [111]. Thus, studies must use larger data samples to improve the results’ generalization and validity. However, collecting large samples of in-context data is a cumbersome task as well. Finally, few works discussed about the challenge of implementing GBA in the classroom, but this is actually 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 [112].

4.3 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 title. 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 have focused our search on Scopus and 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 happening in industry companies and by practitioners. 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 that we are 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 before in [16]. The potential of GBA is now emerging, coming along with the rise of big data. Data mining and visualization techniques on player interaction logs can provide valuable insights to different stakeholders regarding how players interact with the game [113]. The increasing interest in games as a learning method...
also indicates the potential of using them as actual assessment tools. In our review, we found that GBAs are not only being used in formal education but also in medical and professional areas, among others. As expected, the most frequent area where GBAs are being applied is formal 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 applied in many other contexts. Concerning the professional environment, companies have begun to include assessment games for employee recruitment and selection. This is a relatively new trend due to certain limitations, such as the cross-domain generalizability of behaviors between game and workplace contexts, which need further research [51]. 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 living activities, allowing a precise and complete cognitive evaluation, which can be useful to treat certain diseases such as Alzheimer [62] or using games to rehabilitate children with cerebral visual impairment using an eye-tracker [85]. Due to all this, we firmly believe that the future of games for assessment is promising; however, further research is needed to overcome the raised existing problems, increase the still limited application of games in real life environments, in order to start building the classrooms of the future.

5 Conclusions

Technology is changing and improving everyday, and this is also making a significant impact in educational areas. Moreover, playing games is one of the most popular activities over 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 present some challenges, such as finding the time for both the presenter/preceptor and student to learn the systems employed, the financial impacts on both parties, and technical limitations [116]. We can settle 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. The main purpose was to review empirical studies of digital GBAs published until 2020. Based on a detailed systematic review of the 66 selected papers, we concluded that games are mainly used in formal education for assessment and validation purposes. The knowledge 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 of the studies used games designed specifically with 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, being 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 for an actual systematic adoption of games for assessment. For example, the next generation of GBA studies should ensure that enough data is collected in order to have meaningful and reliable results, since one of the main limitations of the current research was 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 forget about 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 results’ reliability and generalization. We also believe that future studies should continue exploring the use of visualizations and dashboards to integrate games in schools, using a more intuitive approach rather than providing teachers with raw numbers outputs or metrics, which are usually harder to understand. Teachers should also have a more important role in future work, in order to address digital and assessment literacy issues, as well as the potential interpretability and actionability of GBAs. Finally, as far as we are concerned, there are no theoretical frameworks within the GBA area (a related one regarding serious games could be [114]). 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 game 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] Lea S Eiland and Timothy J Todd. Considerations when incorporating technology into classroom and experiential teaching. *The Journal of Pediatric Pharmacology and Therapeutics*, 24(4):270–275, 2019.

[2] Sara De Freitas. Are games effective learning tools? a review of educational games. *Journal of Educational Technology & Society*, 21(2):74–84, 2018.

[3] ESA. 2020 essential facts about the computer and video game industry. Technical report, Entertainment Software Association, 2020.

[4] ISFE. Isfe key facts 2020. Technical report, ISFE, 2020.

[5] Stamatios Papadakis. The use of computer games in classroom environment. *International Journal of Teaching and Case Studies*, 9(1):1–25, 2018.

[6] Sebastiaan de Klerk and Pamela M Kato. The future value of serious games for assessment: Where do we go now? *Journal of Applied Testing Technology*, 18(S1):32–37, 2017.

[7] Doug Clow. An overview of learning analytics. *Teaching in Higher Education*, 18(6):683–695, 2013.

[8] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[9] Yoon J Kim and Valerie J Shute. Opportunities and challenges in assessing and supporting creativity in video games. In *Video games and creativity*, pages 99–117. Elsevier, 2015.

[10] Douglas B Clark, Emily E Tanner-Smith, and Stephen S Killingsworth. Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, 86(1):79–122, 2016.

[11] K.E. Dicerbo. Game-based assessment of persistence. *Educational Technology and Society*, 17(1):17–28, 2013.

[12] J.A. Ruiperez-Valiente, M. Gaydos, L. Rosenheck, Y.J. Kim, and E. Klopfer. Patterns of engagement in an educational massively multiplayer online game: A multidimensional view. *IEEE Transactions on Learning Technologies*, 13(4):648–661, 2020.

[13] Yoon Jeon Kim and Dirk Ifenthaler. Game-based assessment: The past ten years and moving forward. In *Game-Based Assessment Revisited*, pages 3–11. Springer, 2019.

[14] Meihua Qian and Karen R Clark. Game-based learning and 21st century skills: A review of recent research. *Computers in human behavior*, 63:50–58, 2016.

[15] Anissa All, Elena Patricia Nunez Castellar, and Jan Van Looy. Measuring effectiveness in digital game-based learning: A methodological review. *International Journal of Serious Games*, 1(2), 2014.

[16] Ming-Chaun Li and Chin-Chung Tsai. Game-based learning in science education: A review of relevant research. *Journal of Science Education and Technology*, 22(6):877–898, 2013.

[17] Cristina Alonso-Fernandez, Antonio Calvo-Morata, Manuel Freire, Ivan Martinez-Ortiz, and Baltasar Fernández-Manjón. Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, 141:103612, 2019.

[18] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjörn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, 2021.

[19] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372, 2021.

[20] José A Ruipérez-Valiente. El proceso de implementación de analíticas de aprendizaje. *RD. Revista Iberoamericana de Educación a Distancia*, 23(2):85–101, 2020.

[21] Areenzoo Aghaie Chadegani, Hadi Salehi, Melor Yunus, Hadi Farhadi, Masood Fooladi, Maryam Farhadi, and Nader Ale Ebrahim. A comparison between two main academic literature collections: Web of science and scopus databases. *Asian social science*, 9(5):18–26, 2013.

[22] Elsevier. About scopus, 2021.

[23] Clarivate. Web of science, 2021.
[24] Alyona Medelyan. Coding qualitative data: How to code qualitative research, 2021.

[25] Manuel J. Gomez, José Ruipérez-Valiente, and Félix 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 December 9, 2021.

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

[27] J.-W. Lin, C.-W. Tsai, and C.-C. Hsu. A comparison of computer-based and game-based formative assessments: a long-term experiment. Interactive Learning Environments, 2020.

[28] J.-P. Lindenmayer, A. Goldring, S. Borne, A. Khan, R.S.E. Keefe, B.J. 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, 223:166–172, 2020.

[29] Stefanie Wiloth, Nele Lemke, Christian Werner, and Klaus Hauer. Validation of a computerized, game-based assessment strategy to measure training effects on motor-cognitive functions in people with dementia. JMIR Serious Games, 4(2):e5696, 2016.

[30] Andrew B Collmus, Michael B Armstrong, and Richard N Landers. Game-thinking within social media to recruit and select job candidates. In Social media in employee selection and recruitment, pages 103–124. Springer, 2016.

[31] E. Short and N. Weidner. Gamers at work: Predicting workplace-relevant behaviours across domains. Journal of Gaming and Virtual Worlds, 11(2):161–177, 2019.

[32] Cătălin Gabriel Ionită Alina Stanciu and Adrian Toscă Dan Florin Stănescu. Development of an integrated game based assessment approach—the next generation of psychometric testing. European Journal of Sustainable Development, 8(5):270–270, 2019.

[33] H.G.K. Hummel, D. Joosten-ten Brinke, R.J. Nadolski, and L.K.J. Baartman. Content validity of game-based assessment: case study of a serious game for ict managers in training. Technology, Pedagogy and Education, 26(2):225–240, 2017.

[34] Agostino Marengo and Alessandro Pagano. Innovative ways to assess soft-skills: the in-basket game online experience. In European Conference on e-Learning, pages 325–XVII. Academic Conferences International Limited, 2020.

[35] M. Ventura and V. Shute. The validity of a game-based assessment of persistence. Computers in Human Behavior, 29(6):2568–2572, 2013.

[36] E.J. Weiner and D.R. Sanchez. Cognitive ability in virtual reality: Validity evidence for vr game-based assessments. International Journal of Selection and Assessment, 28(3):215–235, 2020.

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

[38] R.G. 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), 9505:250–263, 2015.

[39] L.F.Z. Rivera and C.A. Suescún. Game-based assessment for radiofrequency circuits courses in engineering. In Proceedings - Frontiers in Education Conference, FIE 2015, volume 2014, 2015.

[40] 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, 13(4):1075–1083, 2017.

[41] Y.J. Kim, R.G. Almond, and V.J. Shute. Applying evidence-centered design for the development of game-based assessments in physics playground. International Journal of Testing, 16(2):142–163, 2016.

[42] 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, pages 985–991, 2020.

[43] Y. Song and J.R. Sparks. Measuring argumentation skills through a game-enhanced scenario-based assessment. Journal of Educational Computing Research, 56(8):1324–1344, 2019.

[44] Learningrx. What are cognitive skills?, 2021.

[45] Indeed Editorial Team. Cognitive skills: What they are and how to improve them, 2021.
[46] Y. Song and J.R. Sparks. Building a game-enhanced formative assessment to gather evidence about middle school students’ argumentation skills. *Educational Technology Research and Development*, 67(5):1175–1196, 2019.

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

[48] Esraa M Mosalam, Ghada A El Khayat, Shaimaa Lazem, Lilia Cheniti-Belcadh, and Bilal Said. Assessing modelling readiness in a games environment. In *2019 7th International conference on ICT & Accessibility (ICTA)*, pages 1–6. IEEE, 2019.

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

[50] V. Korhonen, H. Räty, and E. Kärnii. 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*, 19(4):281–294, 2017.

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

[52] G.T. 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)*, 10714 LNCS:795–808, 2018.

[53] 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*, 15(3), 2020.

[54] H. Ketamo and K. Devlin. Replacing pisa with global game based assessment. In *Proceedings of the European Conference on Games-based Learning 2014*, volume 1, pages 258–264, 2014.

[55] R. Halverson and V.E. Owen. Game-based assessment: An integrated model for capturing evidence of learning in play. *International Journal of Learning Technology*, 9(2):111–138, 2014.

[56] O. Gaggi, T.M. 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*, 195 LNICST:155–165, 2017.

[57] M. Bertling, G. Tanner Jackson, A. Oranje, and V.E. 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)*, 9112:545–549, 2015.

[58] H.B. Barón, R.G. Crespo, J. Pascual Espada, and O.S. Martínez. Assessment of learning in environments interactive through fuzzy cognitive maps. *Soft Computing*, 19(4):1037–1050, 2015.

[59] 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*, pages 106–110, 2017.

[60] S Pouzezvara, S Powers, G Moore, C Strigel, and K McKnight. Assessing soft skills in youth through digital games. In *ICERI2019 Proceedings*, pages 3057–3066. IATED, 2019.

[61] A. Mavridis and T. Tsiatsos. Game-based assessment: investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, 33(2):137–150, 2017.

[62] V. Vallejo, P. Wyss, L. Rampa, A.V. Mitache, R.M. Müri, U.P. Mosimann, and T. Nef. Evaluation of a novel serious game based assessment tool for patients with alzheimer’s disease. *PLoS ONE*, 12(5), 2017.

[63] A. Leonardou, M. Rigou, and I. Garofalakis. Techniques to motivate learner improvement in game-based assessment. *Information (Switzerland)*, 11(4), 2020.

[64] V.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*, pages 331–342, 2015.

[65] K. Gicali, B.M. 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*, 156, 2020.

[66] 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*, 58(5):1003–1028, 2020.
A Systematic Literature Review of Digital GBA Empirical Studies

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

[68] J.C. 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.

[69] M.C. Gómez-Álvarez, J.A. 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*, 25(4):633–642, 2017.

[70] 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*, 11(3):101–115, 2015.

[71] M. Cutumisu, D.B. Chin, and D.L. 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*, 50(6):2977–3003, 2019.

[72] D.B. Chin, K.P. Blair, and D.L. Schwartz. Got game? a choice-based learning assessment of data literacy and visualization skills. *Technology, Knowledge and Learning*, 21(2):195–210, 2016.

[73] C.M. 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*), 10331 LNAI:502–505, 2017.

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

[75] Luis Tari. *Knowledge Inference*, pages 1074–1078. Springer New York, New York, NY, 2013.

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

[77] R. Levy. Dynamic bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, 54(6):771–794, 2019.

[78] Marshall Hargrave. Deep learning, 2020.

[79] 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*, 30(3):481–503, 2020.

[80] H Michael Chung and Paul Gray. Data mining. *Journal of management information systems*, 16(1):11–16, 1999.

[81] Manuel J Gomez, José A Ruipérez-Valiente, Pedro A Martinez, and Yoon Jeon Kim. Exploring the affordances of sequence mining in educational games. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 648–654, 2020.

[82] Oxford. Oxford languages and google, 2021.

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

[84] 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*, 81:259–269, 2015.

[85] M. Ciman, O. Gaggi, T.M. Sgaramella, L. Nota, M. Bortoluzzi, and L. Pinello. Serious games to support cognitive development in children with cerebral visual impairment. *Mobile Networks and Applications*, 23(6):1703–1714, 2018.

[86] J. Delcker and D. Ifenthaler. Mobile game-based language assessment. *International Journal of Emerging Technologies in Learning*, 15(3):195–206, 2020.

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

[88] J.C. Paiva and J.P. Leal. Asura: A game-based assessment environment for mooshak. In *OpenAccess Series in Informatics*, volume 62, 2018.

[89] C. Harteveld and S.C. Sutherland. The goal of scoring: Exploring the role of game performance in educational games. In *Conference on Human Factors in Computing Systems 2015 - Proceedings*, volume 2015-April, pages 2235–2244, 2015.
A Systematic Literature Review of Digital GBA Empirical Studies

[90] D.J.R. 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*, 55:796–803, 2016.

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

[92] J. Hao and R.J. Mislevy. The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design. *ETS Research Report Series*, 2018(1), 2018.

[93] 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*, pages 727–728, 2016.

[94] 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*, 39(10):1255–1277, 2019.

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

[96] D.R. 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*, 51(1):55–86, 2020.

[97] 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*, 617:126–133, 2017.

[98] Mary Jo Dondlinger. Educational video game design: A review of the literature. *Journal of applied educational technology*, 4(1):21–31, 2007.

[99] James Paul Gee. Are video games good for learning? *Nordic Journal of Digital Literacy*, 1(03):172–183, 2006.

[100] Mete Akcaoglu and Matthew J Koehler. Cognitive outcomes from the game-design and learning (gdl) after-school program. *Computers & Education*, 75:72–81, 2014.

[101] Colin Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[102] Kenneth Cukier. A special report on managing information. *The Economist*, 394(8671):3–18, 2010.

[103] Jeffrey Chin, Richard Dukes, and William Gamson. Assessment in simulation and gaming: A review of the last 40 years. *Simulation & Gaming*, 40(4):553–568, 2009.

[104] Robert J Mislevy and Geneva D Haertel. Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice*, 25(4):6–20, 2006.

[105] Valerie J Shute. Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55(2):503–524, 2011.

[106] Cheryl I. Johnson and Richard E. Mayer. Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, 26(6):1246–1252, 2010. Online Interactivity: Role of Technology in Behavior Change.

[107] Fiona Fidler and John Wilcox. Reproducibility of scientific results. *The Stanford Encyclopedia of Philosophy*, 2018.

[108] Tim van der Zee and Justin Reich. Open education science. *AERA Open*, 4(3):2332858418787466, 2018.

[109] Stuart Buck. Solving reproducibility, 2015.

[110] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.

[111] Giani Petri and Christiane Gresse von Wangenheim. How games for computing education are evaluated? a systematic literature review. *Computers & education*, 107:68–90, 2017.

[112] Manuel J Gomez, José A Ruizpérez-Valiente, Pedro A Martínez, and Yoon Jeon Kim. Applying learning analytics to detect sequences of actions and common errors in a geometry game. *Sensors*, 21(4):1025, 2021.

[113] Manuel Freire, Ángel Serrano-Laguna, Borja Manero, Iván Martínez-Ortiz, Pablo Moreno-Ger, and Baltasar Fernández-Manjón. Game learning analytics: learning analytics for serious games. In *Learning, design, and technology*, pages 1–29. Springer Nature Switzerland AG, 2016.

[114] Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler. Serious games analytics: Theoretical framework. In *Serious games analytics*, pages 3–29. Springer, 2015.