Assessment Measures in Game-based Learning Research: A Systematic Review

Gabriele Gris[1], Clarissa Bengtson[2]
[1] *Graduate Program in Psychology, Federal University of São Carlos, grisgabriele@gmail.com*
[2] *Graduate Program in Education, Federal University of São Carlos clabengtson@gmail.com*

Abstract

The use of games in educational contexts is well documented in GBL research. Nevertheless, effectiveness evidence needs to be more extensively analyzed. An effective GBL strategy should address the learning aspects and promote players' engagement in an easy-to-use system. To gather the information already present in literature, we sought to answer how learning, engagement, and usability of games are evaluated in GBL research. We conducted a systematic review of empirical studies in ERIC, IEEE, Springer, and Web of Science databases. We included 91 studies for the final analysis and categorized their measures and instruments. We find a prevalence of learning assessments over engagement and usability assessments. Learning is mainly evaluated by direct measures, while indirect measures mostly assess engagement and usability. The use of indirect measures and instruments without psychometric qualities compromises the strength of the evidence for the effectiveness of game-based learning. Future studies should add direct assessments and indirect measures with psychometric qualities to assess engagement and usability. The study's limitations are discussed.

Keywords: *Effectiveness, Game-based Learning, Assessment Measures, Learning, Engagement, Usability.*

1 Introduction

For several decades, researchers and professionals have announced multiples ways to use games to improve learning. [1] emphasized that an ideal instructional activity should provide good educational results and emotional satisfaction that rewards learning. The author's central defense is that games are tools with enormous instructional potential.

The use of games to support teaching and learning processes is called game-based learning (GBL [2], [3]), and reports of student acceptance are common [4]. We start from the premise that although the use of games in educational contexts is well documented, the evidence of effectiveness needs to be more extensively studied. We considered that an effective GBL strategy should address the learning aspects and promote players' engagement in an easy-to-use system. Thus, for this study, to assess GBL's effectiveness, we considered the dimensions of learning, engagement, and usability.

1.1 Game-based learning and effectiveness measures

Over the decades of research and application of games as teaching tools, it has become essential to identify effectiveness and map possible contributions in educational contexts. As highlighted by [5], it is necessary to "seek to prove, evidence and solidify the

contributions of technologies, especially digital games, as mediators or enhancers of learning" (p.111, own translation). We can extend this demand to analog games, as well.

There is literature regarding games' effects on learning in different knowledge areas [6]–[10]. Also, reviews have been conducted to assess evidence of game-based learning effectiveness [11]–[15]. Seeking to map the methods used to assess the effectiveness of digital game-based learning (DGBL), [12] conducted a systematic literature review focusing on the research methods adopted. The authors selected 25 studies on digital games with pre and post-test group designs. They collected information about participants, interventions, methods, measures, and results. The authors affirm that it is not possible to generalize GBL's effectiveness due to the diversity of research designs and measures adopted.

From previous reviews [16], [14] conducted a narrative review of the literature. The authors summarized studies on transferring skills learned in digital games for external tasks, improving cognitive processes, playing time and integration with curricular objectives, effects on players, attitudes towards games, cost-effectiveness, and use of games for assessments. Despite the absence of criteria for comparison between studies, the authors state empirical support for games' use for learning purposes. Other studies also revealed positive results of the GBL [17]. The evidence, however, is weaker than the enthusiasm for the use of games suggests. This assertion is supported by methodological reviews of effectiveness in GBL [12], [18]. Besides the difficulty of comparing data between studies, other methodological weakness has been found in GBL research. To answer how games for computing education are evaluated, [18] reviewed 112 articles. 81% did not report well-defined methods to evaluate educational games' impact on learning. In addition, objectives, measures, and instruments were poorly described. [19] conducted a meta-analysis to investigate the effects of learning video games on students' mathematics achievement compared with traditional instructional methods. They analyzed 24 studies and found a small but marginally significant overall impact of the video games' higher learning gains than conventional methods. However, most studies presented incomplete information about video games and GBL interventions.

Despite that, there is great enthusiasm for (D)GBL, mainly due to the typical relationship established between games and engagement [20] [21], [22]. Although some proposals for evaluating educational games considering motivational variables [23], these aspects are not always investigated [12], [13]. Conceptual variations (e.g., engagement and motivation used as synonyms or distinct concepts) can also contribute to GBL assessment's little cohesive literature. The literature about engagement in DGBL research is relatively new and mainly based on psychological theories [11] such as the theory of uses and gratifications [24], [25], the theory of self-determination [26], [27], the ARCS (attention, relevance, confidence, and satisfaction) model [28]–[30], and the concept of flow [31].

Regardless of the relationship between games and engagement, some reviews show that most studies assessed mainly the learning dimension [13], [15]. [13], for example, carried out a systematic review of the literature regarding methods and procedures used to evaluate serious games. They used broad descriptors (evaluation, validation, assessment, serious game, simulation game, education, teaching, and training) and, in general, reported a predominance of the use of questionnaires, in addition to 18 evaluation characteristics. Among these, they highlight the evaluation of teaching objectives.

The reviews by [12] and [13] present different proposals to evaluate games' effectiveness in teaching contexts. The first presents a greater focus on methodological issues, discussing possible implications of the results' validity. The second describes the general characteristics of the studies, such as types of games used. One common point identified in the two reviews is the use of questionnaires to assess games' effects on students' learning and motivation. [11] conducted one of the few reviews that systematically investigated engagement and learning in GBL research. They focus on game design elements and how gameplay engagement may affect learning. However, a detailed analysis of measures used to assess learning and engagement in GBL research falls outside the review's scope.

In addition to the dimensions of learning and engagement, the design's suitability for specific contexts and target audiences should be assessed [32]. The design's adequacy is related to the product's usability and can either function as a facilitator or as an impediment to learning [33]. The International Standards Organization defines usability as the "extent to which a system, product or service can be used by specific users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [34]. The measures used to evaluate these metrics, however, vary broadly [35]. One method used to find usability problems in a user interface design is the heuristic evaluation. An evaluator group usually judges if the interface meets some pre-conceived usability principles. The evaluators point out the system's strongness and weakness and provide recommendations to improve it [36]–[38]. Besides that, it is possible to assess the usability of a system from the user's perspective [39], using the "think aloud" technique [39], [40], or applying instruments such as questionnaires, scales, etc. [35], [41]. A widely used metric to perceived usability is the System Usability Scale (SUS) [35]. Created to assess the usability of industrial systems, the SUS is a ten-items Likert scale and has been applied in several contexts [42], including in GBL research [43], [44].

Although relevant, the usability aspects have been less investigated in GBL research [32]. Engagement and usability issues are strongly related to design elements. Authors such as [45] argue that we should evaluate these dimensions to ensure a serious game's effectiveness.

On this basis, we considered that the dimensions of learning, engagement, and usability are essential to assess GBL effectiveness. Nevertheless, information about the assessment of these three dimensions in GBL context is fragmented. Also, there is little discussion about the use of direct and indirect measures in GBL research to the best of our knowledge. Most of the debate on these measurement methods has been made in the context of summative assessments of learning in Higher Education. In short, a direct measure assesses the knowledge or mastery through the actual work. On the other hand, indirect measures reflect attitudes and opinions, usually obtained from interviews, surveys, and other self-report instruments. Inconsistent results have been reported regarding the correspondence of direct and indirect methods used to assess learning [46]–[48]. We believe that these shreds of evidence can, at least, support reflections on current practices in GBL research.

Given the above, this literature review seeks to combine empirical studies' findings to answer how learning, engagement, and usability aspects of games are assessed in GBL research. A systematic review of measures and instruments used in GBL research may help systematize the field's effectiveness assessment. In this review, learning refers to all possible changes measured regarding the pedagogical outcomes expected. Engagement refers to the probability of continuing playing and how players experience and feel about the game, including flow, motivation, and other phenomena explained by psychological theories. Usability refers to how easy and efficient a system is to use. We deliberately adopt broad characterizations to gather literature with diverse theoretical frameworks and methodological approaches. The measures used in the reviewed studies were categorized based on the three dimensions listed, as well as direct or indirect, seeking to collect the information that, despite the reviews carried out, are still scattered in the literature.

2 Method

This study has been undertaken and reported as a systematic literature review based on PRISMA guidelines [49] (Appendix 1). We used the *StArt* software [50], [51] to manage the search and *Airtable* [52] to store the collected information. In the following sub-sections, we described the activities carried out in each phase of the review.

2.1 Electronic databases and search terms

The electronic databases searched in this review were ERIC, IEEE, Springer, and Web of Science. We chose them, considering that GBL is a multidisciplinary field. We chose search terms from previous systematic reviews [2], [11], [12], [23], [53] and Eric Thesaurus. They focused on the following five categories: (a) games, (b) learning, (c) assessment, (d) engagement, and (e) usability. When necessary, we add terms to exclude proceedings papers and book chapters, focusing on empirical methods.

(a) Games: game, computer game, video game, digital game, gaming, electronic game.

(b) Learning edu* (educational, education, educative), serious, learn* (learning, learner), game based learning, digital game based learning, instruction, classroom.

(c) Assessment: assess* (assess, assessment), evaluat* (evaluation, evaluate), measur* (measure, measurement), effect, impact, outcome, success.

(d) Engagement: engag* (engagement, engage), motivat* (motivation, motivate), enjoy* (enjoy, enjoyment), preference, participation.

(e) Usability: usability, efficiency, effectiv* (effective, effectiveness), satisfaction, difficulty, user friendly, intricacy.

Appendix 2 lists the complete search strings.

2.2 Selection criteria and search procedure

The search focused on empirical studies published between 2013 and 2018 in peer-reviewed journals. Overall, articles were selected if they explore some effectiveness assessment methods (in any dimension studied) and used approaches with game elements. The exclusion criteria identified papers presented as reviews, reports, book chapters, subjects other than assessment measures of games, or did not present game elements in the proposed intervention.

We conducted the systematic review through four phases: identification, screening, eligibility, and inclusion. The first author completed these phases under the supervision of the second author. Thus, we did not calculate the inter-rater agreement because only one author conducted the review's first stages. First, the search strings were applied in each database, and the results were uploaded to StArt Software [50], [51]. Then, in the screening phase, we read titles and abstracts, reviewed them against the inclusion and exclusion criteria, and excluded the irrelevant publications. The remaining articles were advanced to full-text screening. Therefore, we include only publications with relevant information to the study.

We coded the articles that meet the inclusion criteria in two independent data extraction. The agreement between the data extraction was 93,3%. In disagreement cases, we consulted the categories to fit the data better. The protocol gathered information about measures and instruments used to assess learning, engagement, and usability.

2.3 Data analysis

Considering that sometimes the same information was presented in diverse ways, we categorized the measures used in the studies and integrated them with the information already shown in the literature. First, we adopted the categories from the literature consulted before the review. Throughout the study, we also created categories and descriptions to contemplate some measures that were not present in the previously consulted studies.

We also categorized all measures as direct or indirect, based on [46], [54]. We assume that a direct measure requires a demonstration of knowledge or skills and that it provides tangible, visible, and self-explanatory evidence of learning, quality of engagement, or usability. An indirect measure, in contrast, assesses opinions or thoughts about one's own knowledge, skills, learning experiences, perceptions about something, etc. We choose to categorize the results in direct and indirect measures rather than present only the instruments used. A test in a similar structure could provide both types of measures

depending on the context. For example, a multiple-choice mathematics achievement test may offer a direct measure of student's knowledge. A multiple-choice test about perceived usability provided an indirect measure since what we access in this case is the verbal report about perceived usability. Even in one same dimension, these differences may occur. For example, an observation of food consumption is a direct measure, while a report on what was eaten is an indirect measure.

3 Results

First, we present the numerical results (Figure 1) and then analyze the measure categories used to evaluate learning, engagement, usability, and direct and indirect measures.

Initially, we identified 1672 articles, of which 66 were duplicated. Therefore, we read titles and abstracts of 1606 articles, and then, we excluded 1495 records. After the eligibility analysis by reading the eligible full texts, the final sample consisted of 91 articles.

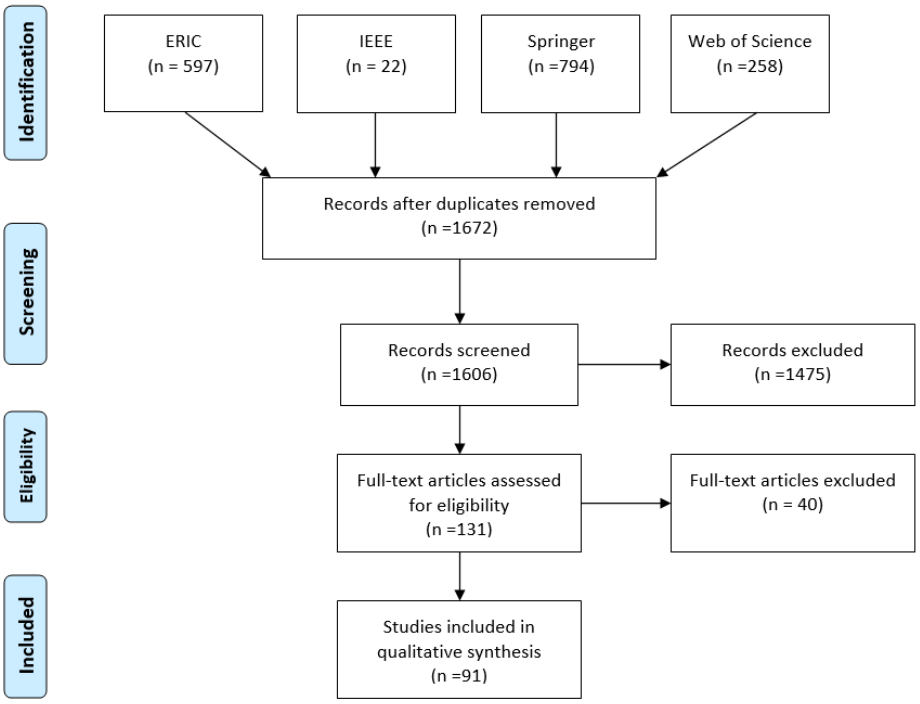

Figure 1. *Flow of information through the different phases of the systematic review*

All studies evaluated learning aspects. 38,4% evaluated only learning aspects, while 31,9% evaluated learning and engagement, 2,2% learning and usability, and 27,5% evaluated all three dimensions.

3.1 Categorization of measures

This section presents all categories of measures found by dimension: learning, engagement, and usability. For each dimension, we present the categories and instruments used. As we mentioned in section 2.3, some categories were based on previous literature, and others were created throughout the coding process. When it is based on literature, the original category and source are presented in the last table's column. Lastly, we presented the measures categorization in direct and indirect, also in accordance with the description in section 2.3.

3.1.1 Learning

More than half of the studies (54%) implemented at least two measures to investigate the learning. Regarding the category of measure, most of the studies (63,7%) implemented tests developed by researchers, and 23 of these (39,6%) only used this type of measure. Only 16,5% of the studies used standardized or literature-reported tests. Less than half of the studies implemented self-efficacy topics (35,2%) and motivation towards educational content (23,1%) as learning measures.

Table 1 gives an overview of the learning assessments, presenting categories, and instruments used to measure this dimension. In some cases, the number of instruments is greater than the frequency of an evaluated category. In some situations, the same category was measured by instruments with more than one format. The same occurs in engagement and usability results.

Table 1. *Measures used to assess learning (N =91 articles)*

Category	N (%)	Instruments	Description	Adapted from
Test scores (literature)	15 (8.24%)	multiple-choice, correct/incorrect assertions, or gaps to complete (4), open-ended questionnaires (4), Likert scale (7), checklists based on observable tasks (2)	Absolute test scores of standardized tests or a test used in previous studies.	Test scores [12]
Test scores (developed for the study)	61 (33.51%)	multiple-choice, correct/incorrect assertions, or gaps to complete (41), open-ended questionnaires (13), not sufficiently described (13)	Absolute test scores of a test developed for the study to evaluate the knowledge about the intervention topic.	Test scores [12]
Student achievement	4 (2.2%)	open-ended questionnaires (2), not sufficiently described (2)	Student achievement in the formal context (e.g., exam scores)	Student achievement [12]
Time on task	5 (2.75%)	the game itself (5)	Time spent on finishing tasks in the game.	Time on the task [12]
Number of errors and correct answers on the game	5 (2.75%)	the game itself (5)	Quantitative data about errors, accurate responses, or the ratio between both.	
Observable behavior changes	3 (1.64%)	checklists based on observable tasks (3) observation (1)	Observable changes in not academic behaviors (e.g., social interaction, food selectivity).	
Self-efficacy topic	24 (13.2%)	Likert scale (22), interview (2), Observation (1)	A verbal description of perceived achievement concerning the instructional topic of the game.	Self-efficacy topic [12]
Sel-efficacy general	6 (3.3%)	Likert scale (6)	A verbal description of academic achievement in general.	Sel-efficacy general [12]

Category	N (%)	Instruments	Description	Adapted from
Perceived educational value	22 (12.08%)	Likert scale (20), interview (1), Observation (1)	The perceived educational value of intervention or knowledge applicability.	Perceived educational value [12]
Teacher expectations	1 (0.55%)	multiple-choice, correct/incorrect assertions, or gaps to complete (1)	Teacher's expectation of change in the students' learning.	Teacher expectations [12]
Motivation towards educational content	36 (19.78%)	Likert scale (35), interview (1)	A verbal description of motivations towards the actual educational content, and not the way it was delivered.	Motivation towards learning/educational content [12]

3.1.2 Engagement

The category "motivation to play and learn" was present in 40 studies, and 14 of these (35%) use this measure exclusively. Table 2 gives an overview of the engagement assessment, presenting the categories and instruments.

Table 2. *Measures used to assess engagement (N =54)*

Category	N (%)	Instruments	Description	Adapted from
Time on game	1 (0.88%)	the game itself (1)	Time spent in the activity, without engaging in parallel tasks or asking to end the game.	Time on the task [55]
Comments about the game.	4 (3.51%)	Likert scale (3), categorized Observation (1)	Comment or suggestion related to gameplay.	Suggestion/Comment [32]
Comments on unrelated subjects	2 (1.75%)	Likert scale (1), categorized Observation (1)	Comments on matters unrelated to the task or game (indicates low engagement).	
Motivation to play and learn	40 (35.09%)	Likert scale (34), open-ended questionnaire (2), interview (1), interview with teachers (2), categorized Observation (3)	Verbal description of the motivation to play and learn.	Motivation towards learning [12], Acceptance, Motivation [13]
Perception of feelings	35 (30.7%)	Likert scale (29), open-ended questionnaire (2), interview (1), multiple-choice questionnaire (1), categorized Observation (1), yes or no questionnaires (1)	Description of feelings while playing (positive or negative).	User experience [13], Satisfied/excited, Pleasantly frustrated, Confuse, Annoyed [32]

Category	N (%)	Instruments	Description	Adapted from
Aesthetic and/or narrative quality	12 (10.53%)	Likert scale (12)	Qualitative assessment of aesthetics or story of the game.	Aesthetic graphics, Fiction/Narrative [45], Content [32]
Immersion/Flow	20 (17.54%)	Likert scale (18), open-ended questionnaire (2)	An intense and immersive concentration that may be related to the distorted perception of time during the game.	

The most assessment was made with Likert scale questionnaires. 41 of 54 (75.93 %) studies that assessed engagement use this type of instrument. Those, only 18 (43.9%), explicitly present its theoretical basis. Seven were based on the ARCS model [29], five on the flow concept [31], [56]–[58], of which one was also based on cognitive load theory [59], and another on the 2x2 achievement goal framework [60]. Additionally, three were established in the Theory of Self-determination [26], one in the Immersion Theory [61], one in the Social-cognitive model of motivation [62], one in the tripartite enjoyment model. One study applies a standardized questionnaire whose theoretical foundation was described during the development, rather than a priori in an expert meeting after two focal groups with players. The experts described the following concepts: competence, flow, suspense, enjoyment, sensory immersion, imaginative immersion, control, negative affect, connectedness, negative affect experience related to playing with others [63]. Thus, 19 of 41 Likert scale questionnaires stated its theoretical basis. It does not mean that more than half of studies apply instruments without any theoretical foundation, only that they do not clearly state it.

Considering the instruments' quality, we verify that six studies used validate instruments based on psychometric properties. The situational motivation scale [64], the Game Immersion Questionnaire [65], Intrinsic Motivation Inventory [26], [66], MAKE framework - based on ARCS model [67], EGame Flow [56], and The Game Experience Questionnaire [63]. All these instruments present at least some evidence of validity and reliability.

In six studies, the researchers adapted validated instruments. Still, in only four, data for internal consistency was presented. 11 studies adapted instruments previously used in the literature but not validated – 10 of which with some internal consistency data. Seven studies developed the instruments used and described the theoretical basis for it. Six of which also presented some data for internal consistency analysis. Finally, eleven questionnaires designed for the studies did not include their base theory, and only four of them offer some internal consistency analysis.

In summary, in only 60.44% of the studies, the engagement dimension was assessed. The most used instruments were Likert without clearly stated theory basis and with low psychometric properties.

3.1.3 Usability

Most of the studies (70.4%) used at least two measures to assess usability. In the 27 studies that investigated usability, we identified 62 evaluations. Table 3 gives an overview of them, presenting the categories and instruments.

Table 3. *Measures used to assess usability (N =27)*

Category	N %	Instruments	Description	Adapted from
Time spent learning how to use the system	2 (3.22%)	Likert scale (1), Observation (1)	The time required to achieve one or more actions that were not previously performed.	Usability [13], Learning [32]

Technical error	3 (4.84%)	Likert scale (1), Observation (1), interview (1)	An unintentional event that prevents the system from functioning correctly.	Technical error [32]
Controls	2 (3.22%)	Likert scale (2)	Adequacy of the controls to the game's actions.	
System responsiveness	3 (4.84%)	Likert scale (2), correct/incorrect assertions (1)	The system's ability to respond continuously to continuously.	Playability [13]
Clarity of tasks	15 (24.2%)	Likert scale (15)	It refers to the presentation of information, data, and facts on the subjects taught.	Content/Information [45]
Ease of use of the system.	17 (27.42%)	Likert scale (16), correct/incorrect assertions (1)	Events related to the intuitive use of the complete system by the user (since before the gameplay started).	Usability [13], Functionality [32]
Clarity of rules	4 (6.45%)	Likert scale (4)	Understanding the rules that describe the possible operations in gameplay.	Mechanics [45]
Clarity of the effects of actions in the game	6 (9.68%)	Likert scale (5), correct/incorrect assertions (1)	Description of the context and consequences of one or more actions performed in the game.	Mechanics [45]
Accessibility	1 (1.61%)	The game itself (1)	Concern to meet the different needs of users.	
Interface	9 (14.52%)	Likert scale (9)	The domain of the means of communication between the user and the system.	Layout/UI [32]

The most used instrument was the Likert scale questionnaires (used in 85.2% of studies). We identify three validated instruments used: EGame Flow *[56]*, System Usability Scale [35], and Attrakdiff2 Scale *[68]*. Lest one was used in a different context of its original validated. Two studies adapted the Instructional Materials Motivation Survey *[69]*, and presented evidence for internal consistency, and one study adapted the Presence Questionnaire *[70]*, without presenting new evidence for internal consistency. Five studies used questionnaires based on the technology acceptance model *[71]*, and present internal consistency evidence.

In summary, the usability dimension was assessed only in 29.67% of the studies analyzed in this review. The most type of instrument used was Likert scales, as well for the engagement dimension. The instruments used present some psychometric properties, although most studies do not use validated instruments.

3.2 Direct and indirect measures

Figure 2 shows that most studies assessed learning mainly by direct or direct and indirect measures combined. Assessments of engagement and usability are conducted almost exclusively by indirect measurements.

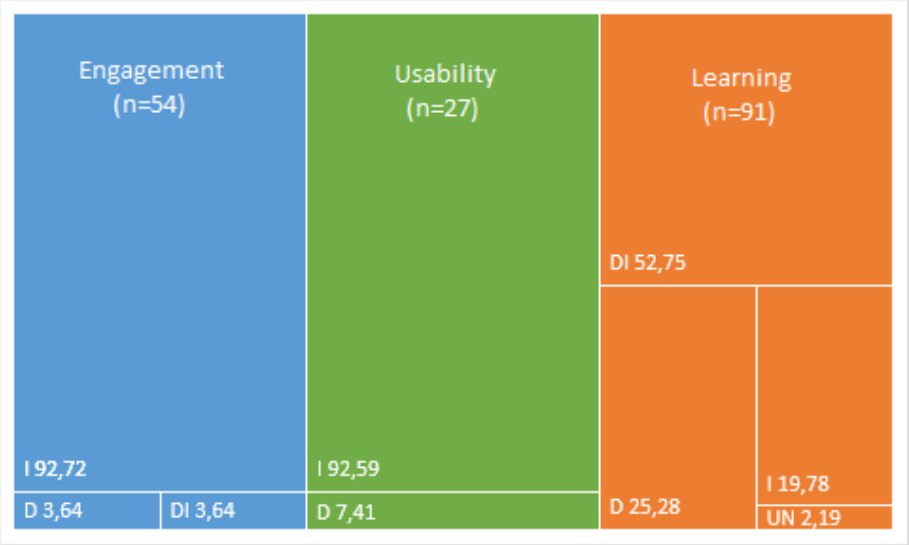

Figure 2. *Frequencies of use of direct and direct measures in the studies reviewed I = Indirect D = Direct DI = Direct and Indirect UN = Unidentified*

4 Discussion

This review focused on identifying how are assessed the learning, the engagement, and the usability of games in GBL research. To discuss these issues, we prepared two sections. In the first part, we discuss the general characteristics and the prevalence of each dimension assessed, confronting our data with the literature. The second section looks specifically at the use of direct and indirect measures in game-based learning assessments.

4.1 Learning, engagement, and usability assessments

Despite the educational potential of GBL, there is a common ground that the empirical evidence is scant [12], [14], [18], [72], [73]. The predominance of learning assessments over engagement and usability assessments, as seen before in the literature [13] - and in this review - can reflect the concern to mainly prove the effectiveness of games in education (and try to change this consensus). The primary interest seems to be in proving that games can teach what they intend to. However, a game is a complex and multifaceted activity. Some authors [69] argued that a scientific attitude regarding all implemented game elements is necessary to take advantage of all benefits offered by games to education and training. Therefore, it is important to conduct assessments beyond the learning dimension.

Our review indicates that the most used measure to assess the engagement is related to the motivation to play and learn and is mostly indirectly investigated. According to [55], there are three methods for gathering information about the player motivation in a GBL environment: through dialog-based communication, game-play-based interaction, and additional equipment. The first one is an indirect measure and the most common in studies. It consists of presenting some questions and asking for a response, a rating, or a self-report, using interviews or questionnaires. The other two methods of gathering information are direct. Through the gameplay-based interaction, it is possible to collect data (e.g., player behavior in the game, task durations, etc.) without interrupting the gameplay. Additional equipment as eye tracker, heart monitor, and others can gather direct measures with minimum gameplay disturbance. In this review, we observed only one occasion in which the game itself recorded the time on the game, and no study used additional equipment to collect data.

The low percentage of studies that evaluated the usability aspects can be related to some differences between games and other software [19], [30]. According to [32, p.11], "the key challenge is that typical usability testing methods focus on measurements that are not necessarily appropriate for games, focusing on aspects such as high productivity, efficacy, and efficiency, as well as low variability, number of errors, and pauses. However, games contemplate reflection, exploration, variety, and trial and error activities". Also, to assess events more related to the game interface and implementation, usability tests should assess some specific events of the user's interaction with the game. These differences may require that researchers and educators create new ways to evaluate or adopt traditional measures, explaining the lack of empirical data in this matter.

To verify if the assessment scenario differs between our sources and conference's proceedings, we took as a sample the articles published in the Games and Learning Alliance conference (GaLA) in the last five years (2016-2020). We choose these proceedings because GaLA is an international conference dedicated to the science and application of serious games and publishes research from different countries. After reading the titles and abstracts of the 210 articles, we applied the same criteria for our original sample and included 132 articles. Then, we randomly selected 30% of the articles and analyzed the assessments conducted. In this small sample (n=39), we observed some differences compared with this study's results. Figure 3 shows the conducted assessments in the studies from both our sample and GaLA proceedings.

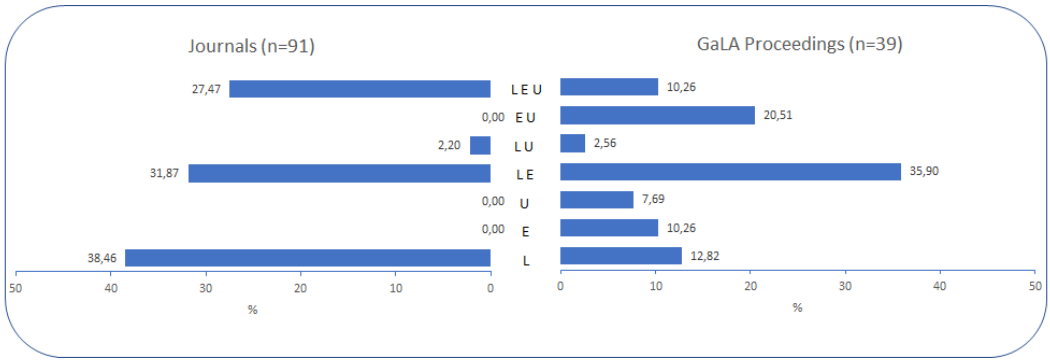

Figure 3. *Dimensions assessed in the studies.*
I = Indirect D = Direct DI = Direct and Indirect UN = Unidentified
L= learning E = Engagement U = Usability

Our results point that while all studies assess the learning, only 60.44% and 29.67% of the sample assessed engagement and usability aspects. In the proceedings, we observed that engagement was the most assessed dimension (in 76.92% of the studies), followed by learning (71.79%) and usability (41.03%). Almost 40% of the conference proceedings (n=14) present preliminary assessments (pilot study, preliminary study, exploratory study, etc.). Another difference refers to the instruments' description. Only 4.05% of the instruments used were not sufficiently described in our original sample, while in the proceedings' sample, this percentage was 21.62%. We also analyzed the use of direct and indirect measures to assess learning, engagement, and usability. Figure 4 shows the use of direct and indirect measures used in the studies from GaLA Proceedings.

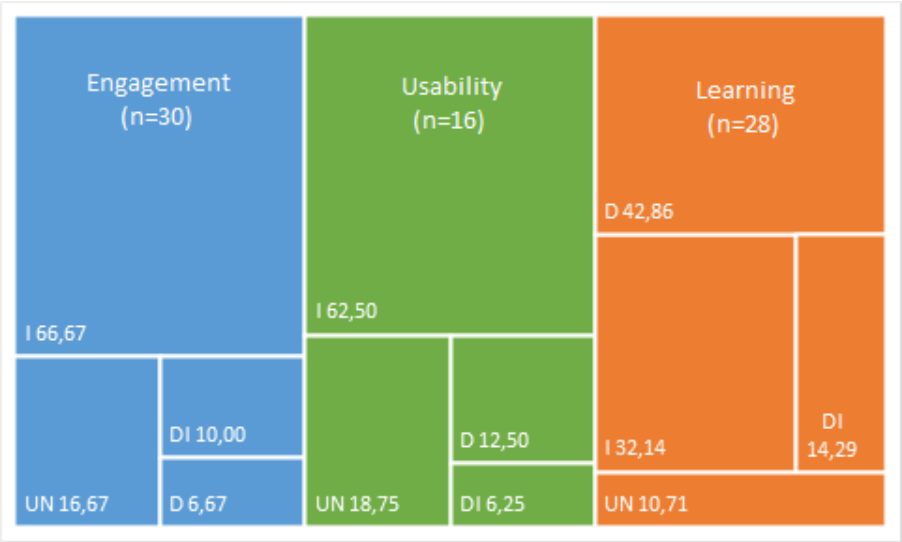

Figure 4. *Frequencies of use of direct and direct measures in the studies from GaLA Proceedings*
I = Indirect D = Direct DI = Direct and Indirect UN = Unidentified

In short, data from GaLA Proceedings shows widely use of indirect measures, primarily to assess engagement and usability, but to a lesser extent of our journal's articles sample. We can infer that usability assessments are performed mainly in the early stages of evaluating serious games (pilot studies, user studies etc.). This would explain the greater frequency of this type of evaluation being found in GaLA Proceedings, since the conference publishes ongoing research. The discussion of ongoing research at scientific events also allows us to improve our studies. On this basis, we point that data from proceedings and journals may differ, and our revision did not outline the complete publication scenario. Even so, we observed some tendencies in both samples: the lower rate in usability assessments compared with the other dimensions and extensive use of indirect measures in GBL research.

Regarding the approach to assessing user interaction usability with the game, there are two ways to gather this data: observational analysis and self-report instruments (surveys, questionnaires, interviews, etc.) [32]. Our review shows a large predominance of data collected from self-report instruments, like what we observe in the engagement assessments. Other studies [11]–[13], [53], and the shreds of evidence from GaLA proceedings present similar results, suggesting that the big picture has not changed in the last years. Also, due to being more common, learning assessments are more diverse. Most studies used direct measures - alone or combined with indirect measures. The learning outcomes are mainly assessed by absolute scores of tests developed for the research or standardized instruments. These data support previous literature [12] , even though this type of assessment is not recommended [11], [12], [48], [73].

4.2 Use of direct and indirect measures

Some hypotheses can explain the prevalence of indirect measures in GBL research. Surveys and questionnaires, for example, are typically easy-to-apply and low-cost tools. Furthermore, there is an understanding that they can offer a viable alternative to assess issues that are difficult to observe systematically [54], [74]. Direct measurement procedures are, indeed, more complex, and time-consuming alternatives. They consist of at least three basics steps: identifying what is to be measured, defining the event in observable terms, and selecting appropriate data-recording procedures to observe, quantify, and summarize all data (frequency, duration, categorization, etc.) [75]. However, we need to cautiously analyze some differences between self-reports (and indirect measures in general) and direct measurement procedures before declaring that they can assess the same events.

[48] presents some evidence regarding indirect measures in a Higher Education context. According to his review, the correlations between longitudinal data and self-reported gains on the same construct (student's growth) are consistently low and often not significantly different from zero. He argues that data from self-reported instruments is problematic and potentially misguiding. Although his data refers to college students' knowledge, we can draw some inferences about self-reports and indirect measures in general.

The main issue with a self-report measure is that what people say about their performance is not always related to how they act [46], [47], [76]. In summary, talking about it is not the same as actually doing something. [77] propose a method for mapping the experience of engagement in video games conceiving intellectual, physical, sensory, social, narrative, and emotional aspects. The authors collected data through reaction cards presented during the gameplay. The participants ought to choose words that best described their motivation to continue playing the game. The authors described the assessment approach as adequate. However, they also assert that the framework does not sufficiently cover the emotional aspect. They inform that the participants' observations - a direct measure - provide evident emotional reactions and recommend video observations to improve the empirical evidence in future works.

Observational approaches are also considered more accurate to assess usability, especially when the aim is to identify specific issues that may prevent unsuccessful interaction with the system [32]. The rationale is the same: it may be more useful to observe the player's behavior and gather objective data from the game itself than to ask (orally or in a questionnaire) what they think or perceive about their experience. Another issue related to self-report as an assessment method implies possible biases such as responding in a socially desirable fashion, agreeing, lying, etc. Of course, direct measures are not entirely unbiased. Observation often requires more than one observer for reliability purposes (e.g., calculating to which extent observers agree), especially when it is impossible to record the interaction. Besides that, when people are aware of being assessed, they can distort both what they say and what they do [76]. These effects may occur even when the participant is tested in the presence of a video recording device [78].

We are not arguing that indirect measures should not be used. However, it is mandatory to know what the instruments do measure and their technical suitability. Research and applied intervention in educational and psychological testing are broad and well-established and may offer GBL research insights. Regarding the tests (questionnaires with Likert scales, checklists, etc.) as of measurement tools, these areas recommend that we look for: updated, detailed, and relevant content, evidence of accurate measurement and reliability with relevant populations, appropriateness of norm, complete technical documentation, proof of validity to support the intended use, and clearly stated limitations [79].

These aspects are related to psychometric concepts and procedures. Reliability, for example, involves the consistency of the tool. A reliable instrument yields the same results when measuring the same thing in identical conditions, considering sources of measurement errors [80], [81]. Moreover, a measurement tool must be valid. Validity implies estimating how well a test measures what it purports. It is assessed based on how well the items cover the content (test content validity). Also, on evaluating the relationship between score obtained in the test and score from others measurement (criterion-related validity), and on executing a critical analysis of how it can be understood within a theoretical framework (construct validity) [81], [82]. The assessment of the degrees of reliability and validity occurs through necessary and adequate statistical procedures. Besides, the process of standardization refers to administrating the test to a representative sample of test-takers to establish normative data. Thus, “a test is said to be standardized when it has clearly specified procedures for administration and scoring, typically including normative data” [83, p. 126].

There are other critical psychometric procedures associated with the development and use of tests. Still, this basic overview of reliability, validity, and standardization shows that methods sought to make indirect measures more accurate and data based. Most tests used

in GBL research – at least among the ones summarized in this review - do not fulfill these requirements. Psychological testing and assessment have a long history [84], while GBL research is a relatively new field. Nevertheless, it is essential to know the instrument's academic and technical suitability to analyze evidence's strength. A standardized tool will provide more solid evidence than an instrument without well-evaluated and documented reliability and validity.

In psychological and educational assessments, both direct and indirect measures are usually applied. Regardless of the type of measure, some steps are essential in these assessments: defining the assessment's objective, choosing the proper instruments, data collection, integration, and interpretation of results [81], [85]. To extrapolate this method to GBL, we first need to define what should be measured. What will indicate that the intervention promotes learning, engagement, or proper usability? We can go further: how do we define each one of these aspects? The definition of a construct is substantial to guide how data may be collected and interpreted. Ideally, the assessments should be developed and adequately analyzed based on a valid theoretical foundation [79], [85]. Only after stating the definition should the measurement tool be selected. A behavior (or trait, or aspect, or phenomena, etc.) may be accessed by more than one measure, direct or indirect [85]. Therefore, a trend in psychological assessments – that may be useful in GBL research is integrating information from varied valid sources.

5 Conclusions

In this study, we identified 91 papers that assess the effectiveness of GBL. We sought to integrate information about learning, engagement, and usability dimensions due to scattered literature. In summary, we found that:

- Learning aspects are much more assessed than engagement and usability features.
- Direct and indirect measures assess learning.
- Indirect measures mainly assess engagement and usability
- Evidence about engagement and usability needs to be carefully analyzed, due to lack of measurement, especially with well-assessed reliability and validity.

In the past decade, GBL's potential was challenged based on insufficient empirical evidence [73]. Although the measures used to assess learning aspects seem to improve over time, we cannot assume the same for engagement and usability. Future works may assess GBL effectiveness based on all three aspects reviewed in this work. Considering that common sense (and almost all GBL supporters) say that games produce great learner engagement and that usability issues may smoothen or prevent learning, the evidence about these aspects should be as strong as possible. Strengths and weaknesses of direct and indirect measures should be considered to design future works in the GBL research context. Therefore, we suggest that future research:

- Seek to access learning, engagement, and usability dimensions.
- Provide operational definitions of learning, engagement, and usability assessments.
- Provide evidence focus on direct measures and/or indirect measures carefully developed with psychometric properties.
- Expand validation studies of instruments to assess all three dimensions.
- Determine the validity of self-reporting measures by comparing participant reports to independently observed data.
- Critically discuss the instruments' cultural appropriateness.

6 Limitations and threats to validity

This literature review has some factors that may have affected its validity. The search terms and databases selected limited the work performed (publication bias). Although we chose the search based on previous literature reviews and educational thesaurus, some words may be missing. The choice to include only articles published in peer-review journals sought to ensure data quality. However, grey literature exclusion probably makes us ignore some relevant works, especially papers published in conference proceedings. Furthermore, the search was conducted by only one researcher. To overcome some of these limitations, all data management was automatized, seeking to reduce human errors.

A challenge faced to gather information in GBL research is the diversity of theoretical and methodological approaches. The definitions are broad, and due to that, the categorization that we made lacks in detail. As a first step, we focus only on analyze assessment measures. However, the complete analysis of the strength of evidence combines these findings to examine methods and designs used in research.

Financial support and data availability

Gabriele Gris was granted a PhD's level scholarship by the Coordination of Superior Level Staff Improvement (CAPES, Project: Applied behavior analysis and assistive technologies for people with cognitive-developmental delay. Covenant 59/2014, number: 88887091031201401). However, the funding source was not involved in study design, in the collection, analysis, and interpretation of data, or the report's writing, and decision to submit the article for publication.

The data that support the findings of this study are openly available in Zenodo repository at doi: 10.5281/zenodo.4383203

Authors' contributions

Both authors contributed to the study conception and design. GG performed material preparation, data collection, and analysis under CB's advisement. GG wrote the first draft of the manuscript, and both authors commented on previous versions. The authors read and approved the final manuscript.

References

[1] C. C. Abt, *Serious games*. Lanham, MD: University Press of America, 1987.

[2] C. Perrotta, G. Featherstone, H. Aston, and E. Houghton, "Game-based learning: Latest evidence and future directions," NFER, Slough, 2013. [Online]. Available: http://ocw.metu.edu.tr/pluginfile.php/10919/mod_resource/content/1/GAME01.pdf.

[3] M. Prensky, *Digital Game-based Learning*. McGraw-Hill, 2001.

[4] M. A. Camilleri and A. Camilleri, "The student's perceptions of digital game-based learning," in *Proceedings of 11th European Conference on Games Based Learning*, Graz, Austria, 2017, pp. 1–14, Accessed: Sep. 20, 2018. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046433.

[5] I. J. Coutinho and L. Alves, "Os desafios e as possibilidades de uma prática baseada em evidências com jogos digitais nos cenários educativos," in *Jogos digitais e aprendizagem: fundamentos para uma prática baseada em evidências.*, 1st ed., I. J. Coutinho and L. Alves, Eds. Campinas: Papirus, 2016, pp. 105–124.

[6] C. Gonzalez, L. D. Saner, and L. Z. Eisenberg, “Learning to stand in the other’s shoes: a computer video game experience of the Israeli–Palestinian conflict,” *Social Science Computer Review*, vol. 31, no. 2, pp. 236–243, Feb. 2013, doi: 10.1177/0894439312453979.

[7] G. Gris, H. W. Alves, G. J. A. Assis, and S. R. Souza, “The use of adapted games for assessment of mathematics and monetary skills,” *Trends in Psychology*, vol. 25, no. 3, pp. 1139–1152, Sep. 2017, doi: 10.9788/tp2017.3-12pt.

[8] H. Pope and C. Mangram, “Wuzzit Trouble: The influence of a digital math game on student number sense,” *International Journal of Serious Games*, vol. 2, no. 4, pp. 5–21, Oct. 2015, doi: https://doi.org/10.17083/ijsg.v2i4.88.

[9] N. Shin, L. M. Sutherland, C. A. Norris, and E. Soloway, “Effects of game technology on elementary student learning in mathematics,” *British Journal of Educational Technology*, vol. 43, no. 4, pp. 540–560, Jul. 2012, doi: 10.1111/j.1467-8535.2011.01197.x.

[10] A. Tripiana-Barbosa and S. R. Souza, “A board game for the teaching, reading, and writing to intellectually disabled people,” *Behavior Analysis: Research and Practice*, vol. 15, no. 1, pp. 90–106, Feb. 2015, doi: 10.1037/h0101073.

[11] A. I. Abdul Jabbar and P. Felicia, “Gameplay engagement and learning in game-based learning: A systematic review,” *Review of Educational Research*, vol. 85, no. 4, pp. 740–779, Dec. 2015, doi: 10.3102/0034654315577210.

[12] A. All, E. P. Nunez Castellar, and J. Van Looy, “Measuring effectiveness in digital game-based learning: a methodological review,” *International Journal of Serious Games*, vol. 2, no. 1, pp. 3–20, Apr. 2014, doi: http://dx.doi.org/10.17083/ijsg.v1i2.18.

[13] A. Calderón and M. Ruiz, “A systematic literature review on serious games evaluation: An application to software project management,” *Computers & Education*, vol. 87, pp. 396–422, Sep. 2015, doi: 10.1016/j.compedu.2015.07.011.

[14] S. Tobias, J. D. Fletcher, and A. P. Wind, “Game-based learning,” in *Handbook od research on educational communications and technology*, 4th ed., J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop, Eds. New York, NY: Springer, 2014, pp. 485–503.

[15] G. Petri and C. G. von Wangenheim, “How to evaluate educational games: A systematic literature review,” *Journal of Universal Computer Science*, vol. 22, no. 7, pp. 992–1021, Jul. 2016, doi: 10.3217/jucs-022-07-0992.

[16] S. Tobias, J. D. Fletcher, D. Y. Dai, and A. P. Wind, “Review of research on computer games,” in *Computer games and learning*, S. Tobias and J. D. Fletcher, Eds. Charlotte, NC: Information Age Publishing, 2011, pp. 127–222.

[17] R. Blunt, “Do serious games work? Results from three studies,” *eLearn*, vol. 2009, no. 12, Dec. 2009, doi: 10.1145/1661377.1661378.

[18] 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, Apr. 2017, doi: 10.1016/j.compedu.2017.01.004.

[19] U. Tokac, E. Novak, and C. G. Thompson, “Effects of game-based learning on students’ mathematics achievement: A meta-analysis,” *Journal of Computer Assisted Learning*, vol. 35, no. 3, pp. 407–420, Jun. 2019, doi: 10.1111/jcal.12347.

[20] B. Karakoç, K. Eryılmaz, E. Turan Özpolat, and İ. Yıldırım, “The Effect of Game-Based Learning on Student Achievement: A Meta-Analysis Study,” *Tech Know Learn*, Sep. 2020, doi: 10.1007/s10758-020-09471-5.

[21] F. Bellotti, B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta, “Assessment in and of serious games: An overview,” *Advances in Human-Computer Interaction*, Article ID 136864, pp. 1–11, Jan. 2013, doi: 10.1155/2013/136864.

[22] R. Van Eck, “Digital Game-Based Learning: It’s not just the digital natives who are restless,” *Educause Review*, vol. 41, no. 2, p. 16, Jan. 2006. [Online]. Available: https://er.educause.edu/articles/2006/1/digital-gamebased-learning-its-not-just-the-digital-natives-who-are-restless

[23] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & Education*, vol. 59, no. 2, pp. 661–686, Sep. 2012, doi: 10.1016/j.compedu.2012.03.004.
[24] W. L. Schramm, J. Lyle, and E. Parker, *Television in the lives of our children*. Stanford, CA: Stanford University Press, 1961.
[25] T. E. Ruggiero, "Uses and gratifications theory in the 21st century," *Mass Communication and Society*, vol. 3, no. 1, pp. 3–37, Feb. 2000, doi: 10.1207/S15327825MCS0301_02.
[26] E. Deci and R. M. Ryan, *Intrinsic motivation and self-determination in human behavior*. Springer US, 1985.
[27] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 54–67, Jan. 2000, doi: 10.1006/ceps.1999.1020.
[28] J. M. Keller, "Development and use of the ARCS Model of instructional design," *Journal of Instructional Development*, vol. 10, no. 3, pp. 2–10, 1987. [Online]. Available: https://www.jstor.org/stable/30221294
[29] J. M. Keller, "The systematic process of motivational design. Performance & Instruction", *Performance & Instruction*, vol. 9-10, nº 26, p. 1–8, Dec. 1987. doi: 10.1002/pfi.4160260902
[30] K. Li and J. M. Keller, "Use of the ARCS model in education: A literature review," *Computers & Education*, vol. 122, pp. 54–62, Jul. 2018, doi: 10.1016/j.compedu.2018.03.019.
[31] M. Csikszentmihalyi, *Flow: The psychology of Optimal Experience*. New York, NY: Harper Perennial Modern Classics, 2008.
[32] P. Moreno-Ger, J. Torrente, Y. G. Hsieh, and W. T. Lester, "Usability testing for serious games: Making informed design decisions with user data," *Advances in Human-Computer Interaction*, vol. 2012, Art nº 69637, pp. 1–13, 2012, doi: http://dx.doi.org/10.1155/2012/369637.
[33] I. R. Perkoski, G. Gris, R. R. Benevides, and S. R. Souza, "Desenvolvimento de jogos educativos com base analítico-comportamental: O procedimento de design iterativo," in *Psicologia e análise do comportamento: Saúde, educação e processos básicos*, J. C. Luzia, G. B. Filgueiras, A. E. Gallo, and J. Gamba, Eds. Londrina: Eduel, 2016, pp. 58–56. [Online]. Available: http://www.uel.br/pos/pgac/wp-content/uploads/2017/03/PSICOLOGIA-E-AN%C3%81LISE-DO-COMPORTAMENTO-SA%C3%9ADE-EDUCA%C3%87%C3%83O-E-PROCESSOS-B%C3%81SICOS.pdf
[34] International Standards Organiozation, "ISO 9241-11:2018(E): Ergonomics of human-system interaction —Part 11:Usability: Definitions and concepts." 2018, Accessed: Nov. 04, 2020. [Online].
[35] J. Brooke, "SUS: 'A quick and dirty' usability scale," in *Usability evaluation in industry*, London, UK: Taylor & Francis, 1996, pp. 189–194.
[36] J. Nielsen, "How to conduct a heuristic evaluation," *Nielsen Norman Group*, 1994. [Online]. Available: https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/ (accessed Nov. 05, 2020).
[37] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Mar. 1990, pp. 249–256, doi: 10.1145/97243.97281.
[38] H. Truong, "Efficient ways to conduct heuristic evaluation for your agile product," *Medium*, Jan. 21, 2020. [Online]. Available: https://uxplanet.org/efficient-ways-to-conduct-heuristic-evaluation-for-your-agile-product-137a4ae4f52d (accessed Nov. 05, 2020).

[39] C.-C. Chang and T. Johnson, "Integrating heuristics and think-aloud approach to evaluate the usability of game-based learning material," *J. Comput. Educ.*, Aug. 2020, doi: 10.1007/s40692-020-00174-5.

[40] J. Nielsen, T. Clemmensen, and C. Yssing, "Getting access to what goes on in people's heads? Reflections on the think-aloud technique," in *Proceedings of the Second Nordic conference on Human-computer interaction*, New York, NY, USA, Oct. 2002, pp. 101–110, doi: 10.1145/572020.572033.

[41] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the System Usability Scale," *International Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, Jul. 2008, doi: 10.1080/10447310802205776.

[42] J. R. Lewis, "The System Usability Scale: Past, present, and future," *International Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, Jul. 2018, doi: 10.1080/10447318.2018.1455307.

[43] S. Perini, R. Luglietti, M. Margoudi, M. Oliveira, and M. Taisch, "Learning and motivational effects of digital game-based learning (DGBL) for manufacturing education –The Life Cycle Assessment (LCA) game," *Computers in Industry*, vol. 102, pp. 40–49, Nov. 2018, doi: 10.1016/j.compind.2018.08.005.

[44] M. M. Marques and L. Pombo, "Game-based mobile learning with augmented reality: Are teachers ready to adopt It?," in *Project and design literacy as cornerstones of smart education*, M. Rehm, J. Saldien, and S. Manca, Eds. Singapore: Springer, 2020, pp. 207–218.

[45] K. Mitgutsch and N. Alvarado, "Purposeful by design: A serious game design assessment framework," in *Proceedings of the International Conference on the Foundations of Digital Games*, New York, NY, 2012, pp. 121–128, doi: 10.1145/2282338.2282364.

[46] C. Luce and J. P. Kirnan, "Using indirect vs. direct measures in the summative assessment of student learning in Higher Education," *Journal of the Scholarship of Teaching and Learning*, vol. 16, no. 4, Art. no. 4, Aug. 2016, doi: 10.14434/josotl.v16i4.19371.

[47] D. R. Bacon, "Comparing direct versus indirect measures of the pedagogical effectiveness of team testing," *Journal of Marketing Education*, vol. 33, no. 3, pp. 348–358, Dec. 2011, doi: 10.1177/0273475311420243.

[48] N. A. Bowman, "Understanding and addressing the challenges of assessing college student growth in student affairs," *Research & Practice in Assessment*, vol. 8, pp. 5–14, Winter, 2013. [Online]. Available: https://eric.ed.gov/?id=EJ1062687

[49] D. Moher, L. Stewart, and P. Shekelle, "Implementing PRISMA-P: recommendations for prospective authors," *Systematic Reviews*, vol. 5, p. 15, Jan. 2016, doi: 10.1186/s13643-016-0191-y.

[50] S. Fabbri, E. Hernandes, A. D. Thommazo, A. Belgamo, A. Zamboni, and C. Silva, "Managing Literature Reviews Information through Visualization," in *Proceedings of the 14th International Conference on Enterprise Information Systems*, Wroclaw, Poland, May 2018, vol. 2, pp. 36–45, doi: 10.5220/0004004000360045.

[51] A. B. Zamboni, A. D. Thommazo, E. C. Hernandes, and Fabbri, S. C. P. F., "StArt: Uma ferramenta computacional de apoio à revisão sistemática.," in *Anais da Brazilian Conference on Software: Theory and Practice - Tools session. UFBA.*, Salvador, 2010, p. 9196, [Online]. Available: http://homes.dcc.ufba.br/~flach/docs/Ferramentas-CBSoft-2010.pdf.

[52] H. Liu, A. Ofstad, and E. Nicholas, *Airtable*. San Francisco, 2012.

[53] E. A. Boyle, T. M. Connolly, T. Hainey, and J. M. Boyle, "Engagement in digital entertainment games: A systematic review," *Computers in Human Behavior*, vol. 28, no. 3, pp. 771–780, May 2012, doi: 10.1016/j.chb.2011.11.020.

[54] M. J. Hansen, "Direct and Indirect Measures of Student Learning," *Planning & Institutional Improvement*, May 22, 2017. [Online]. Available: https://planning.iupui.edu/assessment/prac-files/guidelines/SLMeasures.pdf (accessed May 22, 2020).

[55] I. Ghergulescu and C. H. Muntean, "Measurement and analysis of learner's motivation in game-based e-learning," in *Assessment in Game-Based Learning: Foundations, innovations, and perspectives*, D. Ifenthaler, D. Eseryel, and X. Ge, Eds. New York, NY: Springer, 2012, pp. 355–378.
[56] F.-L. Fu, R.-C. Su, and S.-C. Yu, "EGameFlow: A scale to measure learners' enjoyment of e-learning games," *Computers & Education*, vol. 52, no. 1, pp. 101–112, Jan. 2009, doi: 10.1016/j.compedu.2008.07.004.
[57] J. M. Pearce, M. Ainley, and S. Howard, "The ebb and flow of online learning," *Computers in Human Behavior*, vol. 21, no. 5, pp. 745–771, Sep. 2005, doi: 10.1016/j.chb.2004.02.019.
[58] D. J. Shernoff, M. Csikszentmihalyi, B. Schneider, and E. S. Shernoff, "Student Engagement in High School Classrooms from the Perspective of Flow Theory," in *Applications of Flow in Human Development and Education: The Collected Works of Mihaly Csikszentmihalyi*, M. Csikszentmihalyi, Ed. Dordrecht: Springer Netherlands, 2014, pp. 475–494.
[59] J. Sweller, J. J. G. van Merrienboer, and F. G. W. C. Paas, "Cognitive Architecture and Instructional Design," *Educational Psychology Review*, vol. 10, no. 3, pp. 251–296, Sep. 1998, doi: 10.1023/A:1022193728205.
[60] A. J. Elliot and H. A. McGregor, "A 2x2 achievement goal framework," *Journal of Personality and Social Psychology*, vol. 80, no. 3, pp. 501–519, 2001, doi: 10.1037/0022-3514.80.3.501.
[61] E. Brown and P. Cairns, "A grounded investigation of game immersion," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2004, pp. 1297–1300, doi: 10.1145/985921.986048.
[62] P. R. Pintrich, D. A. F. Smith, T. Garcia, and W. J. Mckeachie, "Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq):," *Educational and Psychological Measurement*, vol. 3, no. 53, pp. 801–813, 1993, doi: 10.1177/0013164493053003024.
[63] K. Poels, Y. A. W. de Kort, and W. A. IJsselsteijn, *D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games*. Eindhoven: TU Eindhoven, 2007. [Online]. Available: https://pure.tue.nl/ws/files/21666952/Fuga_d3.3.pdf
[64] F. Guay, R. J. Vallerand, and C. Blanchard, "On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS)," *Motivation and Emotion*, vol. 24, no. 3, pp. 175–213, Sep. 2000, doi: 10.1023/A:1005614228250.
[65] M.-T. Cheng, H.-C. She, and L. A. Annetta, "Game immersion experience: its hierarchical structure and impact on game-based science learning," *Journal of Computer Assisted Learning*, vol. 31, no. 3, pp. 232–253, Apr. 2015, doi: https://doi.org/10.1111/jcal.12066.
[66] R. M. Ryan, "Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory," *Journal of Personality and Social Psychology*, vol. 43, no. 3, pp. 450–461, Sep. 1982, doi: 10.1037/0022-3514.43.3.450.
[67] H. Haruna, X. Hu, S. K. W. Chu, and R. R. Mellecker, "Initial Validation of the MAKE Framework: A Comprehensive Instrument for Evaluating the Efficacy of Game-Based Learning and Gamification in Adolescent Sexual Health Literacy," *Annals of Global Health*, vol. 85, no. 1, Art. no. 1, Feb. 2019, doi: 10.5334/aogh.1110.
[68] M. Hassenzahl, M. Burmester, and F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität," in *Mensch & Computer 2003: Interaktion in Bewegung*, G. Szwillus and J. Ziegler, Eds. Wiesbaden: Vieweg+Teubner Verlag, 2003, pp. 187–196.
[69] J. M. Keller, *IMMS: Instructional materials motivation survey*. Tallahassee, FL: Florida State University, 1987.

[70] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments: A Presence Questionnaire," *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 3, pp. 225–240, Jun. 1998, doi: 10.1162/105474698565686.
[71] F.-Y. Pai and K.-I. Huang, "Applying the Technology Acceptance Model to the introduction of healthcare information systems," *Technological Forecasting and Social Change*, vol. 78, no. 4, pp. 650–660, May 2011, doi: 10.1016/j.techfore.2010.11.007.
[72] D. Ifenthaler, D. Eseryel, and X. Ge, "Assessment for Game-Based Learning," in *Assessment in Game-Based Learning: Foundations, innovations, and perspectives*, D. Ifenthaler, D. Eseryel, and X. Ge, Eds. New York, NY: Springer, 2012, pp. 1–8.
[73] R. E. Clark, "Learning from Serious Games? Arguments, evidence, and research suggestions," *Educational Technology*, vol. 47, no. 3, pp. 56–59, Jun. 2007 [Online]. Available: https://www.jstor.org/stable/44429512
[74] T. W. Banta and C. A. Palomba, *Assessment essentials: Planning, implementing, and improving assessment in Higher Education*, 2nd ed. San Francisco, CA: Jossey Bass, 2014.
[75] D. L. Gast, "General factors in measurement and evaluation," in *Single case research methodology: Applications in Special Education and Behavioral Sciences*, 2nd ed., D. L. Gast and J. R. Ledford, Eds. New York, NY: Routledge, 2014, pp. 85–104.
[76] A. E. Kazdin, *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press, 1982.
[77] H. Schønau-Fog and T. Bjørner, "'Sure, I Would Like to Continue': A Method for mapping the experience of engagement in video games," *Bulletin of Science, Technology & Society*, vol. 32, no. 5, Special Issue: Game On: The Challenges and Benefits of Video Games, Part I, pp. 405–412, Oct. 2012, doi: 10.1177/0270467612469068.
[78] M. Constantinou, L. Ashendorf, and R. J. McCaffrey, "Effects of a third party observer during neuropsychological assessment," *Journal of Forensic Neuropsychology*, vol. 4, no. 2, pp. 39–47, Jul. 2005, doi: 10.1300/J151v04n02_04.
[79] International Test Commission, "International guidelines for test use," *International Journal of Testing*, vol. 1, no. 2, pp. 93–114, 2001. [Online]. Available: https://www.intestcom.org/files/guideline_test_use.pdf
[80] R. J. Cohen, M. E. Swerdlik, and E. D. Sturman, "Reliability," in *Psychological testing and assessment: an introduction to tests and measurements*, 9th ed., New York, NY: McGraw-Hill Education, 2018, pp. 141–174.
[81] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 2014.
[82] R. J. Cohen, M. E. Swerdlik, and E. D. Sturman, "Validity," in *Psychological testing and assessment: an introduction to tests and measurements*, 9th ed., New York, NY: McGraw-Hill Education, 2018, pp. 175–199.
[83] R. J. Cohen, M. E. Swerdlik, and E. D. Sturman, "Of tests and testing," in *Psychological testing and assessment: an introduction to tests and measurements*, 9th ed., New York, NY: McGraw-Hill Education, 2018, pp. 115–140.
[84] R. J. Cohen, M. E. Swerdlik, and E. D. Sturman, "Historical, cultural, and legal/ethical considerations," in *Psychological testing and assessment: an introduction to tests and measurements*, 9th ed., New York, NY: McGraw-Hill Education, 2018, pp. 36–74.
[85] C. T. Reppold and L. G. Gurgel, "O papel do teste na avaliação psicológica," in *Psicometria*, 1st ed., C. S. Hutz, D. R. Bandeira, and C. M. Trentini, Eds. Porto Alegre: Artmed, 2015, pp. 147–164.

Appendix 1: PRISMA-P 2015 Checklist

This checklist has been adapted for use with protocol submissions to Systematic Reviews from Table 3 in Moher D et al: Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews 2015 4:1

The column "comments" was added by the researchers.

Section/topic	#	Checklist item	Information reported		Line number(s)	Comments
			Yes	No		
ADMINISTRATIVE INFORMATION						
Title						
Identification	1a	Identify the report as a protocol of a systematic review	☒	☐	1-2 (p.3)	
Update	1b	If the protocol is for an update of a previous systematic review, identify as such	☐	☒		Does not apply
Registration	2	If registered, provide the name of the registry (e.g., PROSPERO) and registration number in the Abstract	☐	☒		
Authors						
Contact	3a	Provide name, institutional affiliation, and e-mail address of all protocol authors; provide physical mailing address of corresponding author	☒	☐	3-7 (p.3)	
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review	☒	☐	25-28 (p.17)	
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments	☐	☒		Does not apply
Support						
Sources	5a	Indicate sources of financial or other support for the review	☒	☐	16-19 (p.17)	
Sponsor	5b	Provide name for the review funder and/or sponsor	☒	☐	1-17 (p.17)	
Role of sponsor/funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol	☒	☐	19-21 (p.17)	
INTRODUCTION						
Rationale	6	Describe the rationale for the review in the context of what is already known	☒	☐	42-44 (p.3) 1-55 (p.4) 1-31 (p.5)	

Section/topic	#	Checklist item	Information reported		Line number(s)	Comments
			Yes	No		
Objectives	7	Provide an explicit statement of the question(s) the review will address	☒	☐	32-34 (p.5)	
METHODS						
Eligibility criteria	8	Specify the study characteristics (e.g., PICO, study design, setting, time frame) and report characteristics (e.g., years considered, language, publication status) to be used as criteria for eligibility for the review	☒	☐	22-27 (p.)	
Information sources	9	Describe all intended information sources (e.g., electronic databases, contact with study authors, trial registers, or other grey literature sources) with planned dates of coverage	☒	☐	2-3 (p.6)	
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated	☒	☐	3-17 (p.6)	Including the appendix cited.
STUDY RECORDS						
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review	☒	☐	46-48 (p.5)	
Selection process	11b	State the process that will be used for selecting studies (e.g., two independent reviewers) through each phase of the review (i.e., screening, eligibility, and inclusion in meta-analysis)	☒	☐	26-34 (p.6)	
Data collection process	11c	Describe planned method of extracting data from reports (e.g., piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators	☒	☐	35-37 (p.6)	
Data items	12	List and define all variables for which data will be sought (e.g., PICO items, funding sources), any pre-planned data assumptions and simplifications	☒	☐	37-38 (p.6)	
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale	☒	☐	41-52 (p.6)	
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including	☐	☒		The review itself presents a

Section/topic	#	Checklist item	Information reported		Line number(s)	Comments
			Yes	No		
		whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis				critical analysis of the studies included.
DATA						
Synthesis	15a	Describe criteria under which study data will be quantitatively synthesized	☐	☒		Does not apply.
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data, and methods of combining data from studies, including any planned exploration of consistency	☐	☒		Does not apply.
	15c	Describe any proposed additional analyses (e.g., sensitivity or subgroup analyses, meta-regression)	☐	☒		Does not apply.
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned	☒	☐	21-27 (p.7)	
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (e.g., publication bias across studies, selective reporting within studies)	☒	☐	11-20 (p.13) 2-14 (p.17)	
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (e.g., GRADE)	☐	☒		The review itself presents a critical

Appendix 2: Search strings

Table S1 - *Search strings used.*

Database	String
ERIC	(game OR (computer game) OR (video game) OR (digital game) OR gaming OR (electronic game)) AND (educational OR education OR educative OR serious OR learning OR learner OR (game based learning) OR (digital game based learning) OR instruction OR classroom OR academic) AND (assessment OR asses OR evaluate OR evaluation OR measure OR measurement OR effect OR impact OR outcome OR success OR evidence) AND (engagement OR engage OR motivation OR motivate OR enjoy OR enjoyment OR preference OR participation) AND (usability OR efficiency OR effective OR effectiveness OR (user friendly) OR satisfaction OR difficulty OR intricacy) AND (experimental OR quasi-experimental OR quasiexperimental OR empirical)
IEEE	(game OR "computer game" OR "video game" OR "digital game" OR gaming OR "electronic game") AND (educational OR education OR educative OR serious OR learning OR learner OR "game-based learning" OR "digital game-based learning" OR instruction OR classroom OR academic) AND (assessment OR assess OR evaluate OR evaluation OR measure OR measurement OR effect OR impact OR outcome OR success OR evidence) AND (engagement OR engage OR motivation OR motivate OR enjoy OR enjoyment OR preference OR participation) AND (experimental OR quasiexperimental OR quasi-experimental OR empirical)
Springer	(game OR (computer game) OR (video game) OR (digital game) OR gaming OR (electronic game)) AND (educational OR education OR educative OR serious OR learning OR learner OR (game-based learning) OR (digital game-based learning) OR instruction OR classroom OR academic) AND (assessment OR asses OR evaluate OR evaluation OR measure OR measurement OR effect OR impact OR outcome OR success OR evidence) AND (engagement OR engage OR motivation OR motivate OR enjoy OR enjoyment OR preference OR participation) AND (usability OR efficiency OR effective OR effectiveness OR (user friendly) OR satisfaction OR difficulty OR intricacy) AND (experimental OR quasi-experimental OR quasiexperimental OR empirical) AND NOT (annals OR meeting OR proceedings OR congress OR conference OR chapter OR encyclopedia OR book OR report OR handbook)
Web of Science	((((ALL=((game OR "computer game" OR "video game" OR "digital game" OR gaming OR "electronic game") AND (edu* OR serious OR learn* OR "game base learning" OR "digital game based learning" OR instruction OR classroom OR academic) AND (assess* OR evaluat* OR measur* OR effect OR impact OR outcome OR success OR evidence) AND (engag* OR motivat* OR enjoy* OR preference OR participation) AND (usability OR efficiency OR effectiv* OR satisfaction OR "user friendly" OR difficulty OR intricacy) AND (experimental OR quasiexperimental OR quasi-experimental OR empirical))))))

Note. The reviewers warn us of the terms "game based learning" and "digital game based learning" used without the "-" in the search strings. We then conducted a new search to compare the results using strings with "game-based learning" and "digital game-based learning". In three databases (IEEE, Springer, and Web of Science), the results were identical. In ERIC database, we identified more results without the "-". Therefore, we assume that the absence of the "-" in the strings search does not undermine the review.