


An Analysis of the Evaluation Methods being Applied to Serious Games for Autistic Children

Ana Paula de Carvalho  [Universidade Federal de Minas Gerais | ana.carvalho@dcc.ufmg.br]

Camila Santana Braz  [Universidade Federal de Minas Gerais | camilabraz@ufmg.br]

Raquel Oliveira Prates  [Universidade Federal de Minas Gerais | rprates@dcc.ufmg.br]

✉ *Computing Science Department, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, 31270-901, Brazil.*

Received: 28 March 2023 • **Accepted:** 10 November 2023 • **Published:** 01 January 2024

Abstract Autism Spectrum Disorder is a neurodevelopment condition that significantly impacts social communication and interaction as well as behavior impairments, including restricted and repetitive patterns of behavior, interests, or activities. In recent years, numerous studies have proposed serious games as a way to aid in the therapy of children with ASD. Hence, it is crucial to evaluate the effectiveness of such games and obtain robust evidence of their positive influence on this type of treatment. In this study, we aim to explore the evaluation of games for autistic children by conducting a Systematic Literature Review. We analyze the methods utilized to evaluate these games, their application and combination, the quality aspects assessed, and the number and characteristics (e.g., age and special need) of the participants involved in the evaluation process. Furthermore, we present a compilation of the study findings for each evaluation method. Our findings reveal that there is no standardized methodology since different methods have been utilized and combined in various ways to evaluate serious games that support the treatment of ASD children. As contributions, this paper provides valuable insights into how serious games have been evaluated in this context and can be useful for researchers and game designers working in the field.

Keywords: Autism spectrum disorder, Serious game, Evaluation methods

1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized mainly by persistent deficits in social communication and social interaction, along with repetitive and restrictive patterns of behavior, interests, or activities. According to studies by the Centers for Disease Control and Prevention of the United States, conducted on 8-year-old children in 2018, the prevalence of ASD is 1 in every 44 children [Maenner *et al.*, 2021]. In Brazil, there are no official studies that estimate the prevalence of this disorder in the country [Sukiennik *et al.*, 2021].

In recent years, different studies have been investigating technologies to support the treatment of individuals with ASD [Cordeiro *et al.*, 2018; Virnes *et al.*, 2015]. In this sense, different types of technologies and solutions have been explored. For example, Jouaiti and Henaff [2019] investigated robot-based solutions to motor rehabilitation for children with ASD. Khowaja *et al.* [2020] examined the use of augmented reality to improve various skills of children and adolescents with ASD. In their turn, Glaser and Schmidt [2021] investigated virtual reality intervention design patterns for individuals with ASD. Additionally, multiple researchers have explored the use of serious games, with different focuses, aimed at individuals with ASD (e.g., Xianmei [2017]; Hassan *et al.* [2021]; Kirst *et al.* [2022]; Parisa Ghanouni and Lucyshyn [2021]; Carreño-León *et al.* [2021]; Vallefucio *et al.* [2021]). A game can be defined as a “Serious Game” if its purpose is not only to entertain the player but also to develop a skill [Ritterfeld *et al.*, 2009; Susi *et al.*, 2007].

Serious games are a type of technological solution that has

been widely investigated in order to support the treatment of ASD [Tsikinas and Xinogalos, 2019; Hassan *et al.*, 2021]. Consequently, it is crucial that such games are evaluated to obtain solid evidence of their positive impact on this type of treatment.

Thus, this research aims to investigate how serious games for children with Autism Spectrum Disorder are being evaluated. To achieve this goal, we conducted a Systematic Literature Review (SLR). This work is an extended and revised version of a previous conference paper [de Carvalho *et al.*, 2022], that analyzed which evaluation methods were used, how these methods were applied and combined, what quality properties were evaluated, and the number and profile of the participants involved in the different evaluation types. In addition to the results of these analyses, this paper includes a characterization of the stakeholders involved in the evaluations and the presentation of a compilation of the study findings for each evaluation method.

The results presented and discussed in this work contribute to: 1) advance the knowledge about how serious games aimed at children with ASD have been evaluated, and 2) support serious game researchers and developers in understanding aspects considered in evaluating games and making choices on how to conduct evaluations in their own projects.

The article is organized as follows: Section 2 provides an ASD overview. The related works on the research are discussed in Section 3. The methodology adopted to conduct the literature review and respond the research questions are detailed in Section 4. Section 5 presents and discusses the results of our analysis, and Section 6 presents a characteriza-

tion of the use of the evaluation methods. Section 7 addresses the threats to the validity of the results of this research. Finally, in Section 8, some final considerations are presented.

2 Autism Spectrum Disorder

This section presents the main characteristics of Autism Spectrum Disorder, based on the Diagnostic and Statistical Manual of Mental Disorders - DSM-5 [American Psychiatric Association, 2013].

Neurodevelopmental disorders are conditions that typically manifest in early childhood. Deficits in children's development can harm their personal, social, academic, or professional functioning. Individuals often present the co-occurrence of more than one neurodevelopmental disorder. Neurodevelopmental disorders include intellectual disabilities, communication disorders, autism spectrum disorder, attention-deficit/hyperactivity disorder, specific learning disorder, and motor neurodevelopmental disorders [American Psychiatric Association, 2013].

The essential characteristics of ASD are persistent communication and social interaction deficits, accompanied by repetitive and restrictive patterns of behavior, interests, or activities. Currently, ASD encompasses disorders previously called early infantile autism, infantile autism, Kanner autism, high-functioning autism, atypical autism, unspecified developmental disorder, childhood disintegrative disorder, and Asperger's disorder [American Psychiatric Association, 2013].

The losses in communication and social interaction in multiple contexts include deficits in social and emotional reciprocity (i.e., the ability to engage and share ideas and feelings with others), deficits in nonverbal communicative behaviors used for social interaction, and deficits in the development, maintenance, and understanding of relationships. The repetitive and restricted patterns include simple motor stereotypes (e.g., waving hands), repetitive use of objects (e.g., lining up objects), repetitive speech (e.g., echolalia), insistence on certain routines (e.g., consuming the same foods daily), ritualized verbal or nonverbal behavioral patterns (e.g., traversing a perimeter), restricted and hyper-focused interests, and hypersensitivity/hyposensitivity to sensory stimuli (e.g., no reaction to pain, fascination with rotating objects or lights) [American Psychiatric Association, 2013].

The term "spectrum" in autism refers to the fact that the disorder presents itself in a wide range of symptoms, which vary according to the severity of the autistic condition, the level of development, and the age of the individual. The severity levels for autism are: level 1 ("Requiring support"), level 2 (Requiring substantial support"), and level 3 ("Requiring very substantial support"). These levels are based on the amount of support required by individuals due to impairments in social communication and in restrictive and repetitive patterns of behavior [American Psychiatric Association, 2013].

Individuals with autism also commonly present with other conditions such as language impairment, intellectual impairment, specific learning difficulties (reading, writing, and arithmetic), developmental coordination disorder, and restrictive/avoidant feeding disorder. The diagnosis of autism

in boys is four times higher than in girls [American Psychiatric Association, 2013].

3 Related work

Our research focuses on serious games aimed at children with ASD. In this context, different studies have investigated various aspects of these games. For example, Noor *et al.* [2012] conducted a systematic review and presented an overview of serious games for children with ASD, focusing on the purpose of the game, its type, and the technologies used to develop it. Zakari *et al.* [2014] classified serious games for autistic children with respect to the technological platform, the purpose of the game, the type of graphics (i.e., 2D or 3D), game aspects, and user interaction devices. Tsikinas *et al.* [2016] classified serious games for people with an intellectual disability or ASD based on adaptive behavior and intellectual functioning skills that the games aim to develop and their potential effects. Meanwhile, Xianmei [2017] presented an overview of somatic games (i.e., video games operated by body movements) aimed at autistic children, focusing on the game features, implementation of interventions, and their effectiveness. Kousar *et al.* [2019] presented a comparison of serious games for autistic children, focusing on the purpose of the game, type of autism, technological platform, age, type of graphics (i.e., 2D or 3D) and category. Tsikinas and Xinogalos [2019] studied the effects of computer serious games on people with an intellectual disability or ASD. Hassan *et al.* [2021] evaluated the design of serious games aimed at improving the social and emotional intelligence of children with ASD. Silva *et al.* [2021] compared the use of serious games and entertainment games in interventions for treating ASD.

Although these works addressed serious games for children with autism, only some of them [Tsikinas *et al.*, 2016; Tsikinas and Xinogalos, 2019; Hassan *et al.*, 2021] investigated aspects related to game evaluation. On the other hand, in the literature, different literature reviews explored aspects related to the evaluation of serious games. For example, Calderón and Ruiz [2015] conducted a systematic review to investigate the state of the art of procedures, techniques, and methods used to evaluate serious games in different domains. In addition, the authors analyzed the specific context of evaluating serious games aimed at the software project management area. Yanez-Gomez *et al.* [2017] conducted a systematic review on usability evaluation in serious games. Petri and Gresse von Wangenheim [2017] investigated, in their literature review, how games aimed at teaching computing are evaluated. In their turn, Marques and Monte [2022] conducted a systematic mapping of the literature to investigate how software technologies are being evaluated with users with autism.

Even though there are various studies analyzing serious games and technologies for individuals with ASD or evaluating serious games, none of them have specifically focused on the evaluation of serious games for children with ASD. The few that have addressed the subject [Tsikinas *et al.*, 2016; Tsikinas and Xinogalos, 2019; Hassan *et al.*, 2021] did so tangentially amidst other issues. Therefore, this study has a

more specific focus and extends the contributions of these works to the evaluation of serious games for children with ASD. As a differential, we present an overview of the different methodologies employed in the studies; we map these methodologies to the category of the evaluated game; we describe how they have been combined and applied, and describe the profile and number of participants in the user evaluations and finally, the criteria being used to evaluate the games' quality.

4 Methodology

This study followed the guidelines for conducting a SLR indicated by Kitchenham [2004]. In this section, we present the methodology conducted in our SLR.

4.1 Research questions

Since the aim of this work is to analyze the methods being applied in evaluating serious games aimed at children with ASD, the following research questions were formulated:

RQ1: What evaluation methods have been used?

RQ2: What methods have been used to evaluate each game category?

RQ3: How have the methods been applied in evaluations?

RQ4: What is the sample size and profile of the participants in the evaluations?

RQ5: What quality properties have been evaluated in the studies?

4.2 Search process

We know that the topic addressed in the SLR is multidisciplinary and can be addressed in different areas of knowledge, such as education and health. However, this work focuses on analyzing how serious games have been evaluated from the perspective of the Human-Computer Interaction (HCI) area to help in the future design and evaluation of games. For this reason, the searches were conducted in some of the main repositories and proceedings of events that store relevant works in Computer Science related to HCI and serious games: IEEE Xplore, ACM Digital Library, Entertainment Computing, SBGames (Brazilian Symposium on Games and Digital Entertainment), SBSC (Brazilian Symposium on Collaborative Systems), SBIE (Brazilian Symposium on Informatics in Education), and JIS (Journal on Interactive Systems)¹. Since the ACM and IEEE libraries allow for automated searches with the application of filters, we defined a search string to select the publications from these databases.

It is worth noting that the scope of our research was the analysis of serious games aimed at children with autism spectrum disorder, including other aspects besides their evaluation. Thus, we defined a more comprehensive *search string* to collect studies related to this context².

¹As IHC (Brazilian Symposium on Human Factors in Computing Systems) is published in the ACM Digital Library, it was not necessary to analyze it separately.

²As a result of our broader research, other analyses have been conducted and presented in other works [de Carvalho *et al.*, 2023]

The search *string* considered was: (*autism*³ AND *children AND game*). Subsequently, to conduct the analyses presented in this work, we filtered only those that presented a game's evaluation process from the selected studies.

The SLR was conducted following these steps: (1) Initial search; (2) Elimination by title and abstract; (3) Elimination by diagonal reading (i.e., introduction and conclusion); and (4) Complete reading and data extraction. These steps were conducted following inclusion, exclusion, and quality criteria. In each step, the researchers recorded the data of interest in control spreadsheets.

In the *initial search*, manual searches conducted in the proceedings of events and repositories that did not allow automatic search included all available publications in initial set of articles. In turn, automated searches executed in digital libraries that allowed the application of filters in the search resulted in the selection of works that included the terms of the search *string* in the article content.

The following inclusion criteria were applied: studies from January 2010 to March 2020 that presented specific serious games for children with autism or that included this population (e.g., games aimed at people with neurodevelopmental disorders). When more than one study focused on the same game, the most complete study was considered, and the others were discarded. As for the exclusion criteria, studies not written in English or Portuguese were eliminated. Robots-related studies were also eliminated, as they were outside the scope of our research.

For this study, for the complete reading step (step 4), the following quality criteria were considered: the work must clearly present its goal, its research questions, its methodology, its results and contributions, and present the evaluation process of the games. Articles that did not meet these criteria were excluded.

It is important to note that some precautions were taken to minimize interpretation biases during the elimination and selection of articles during the SLR. In each stage, two researchers analyzed the works, and then a consolidation analysis was carried out between the two. Additionally, if there were any doubts about whether an article should be eliminated, it was kept to be analyzed in the next stage. In the last step, in case of doubts regarding a particular article, a third researcher participated in the discussion and decision-making.

4.3 Data extraction

The following data was extracted from the selected studies: article title, reference, the skill(s) the game aims to improve, target audience, game name, sample size, profile and age range of the evaluation participants, methods used in the evaluation, and criteria used to evaluate⁴.

In the initial search, the automatic search in the IEEE and ACM libraries returned 479 articles. From the other repositories and event proceedings that did not have this mechanism,

³Tests were carried out, which showed that the word *autism* appeared in articles that also used other terms to refer to TEA. For this reason, we only used this term.

⁴The extracted data is available at <https://docs.google.com/spreadsheets/d/14u3skqIRQvdUA0oABWxFciH1keq8ACoSjLJdPqSQTu0/edit?usp=sharing>

all available studies were initially selected, resulting in 3522 works. In the elimination by reading title and abstract (step 2), 4001 articles were analyzed and 239 passed to the next step. In the elimination by diagonal reading step, 162 studies passed to the final step. Finally, after the complete reading and evaluation of the quality of the articles, 70 were considered relevant for our analysis.

In three of these articles, more than one game was presented and evaluated. In addition, only one game was the subject of two articles (but as one presented the game and a preliminary evaluation, and the other focused on a more detailed evaluation of the game, both were kept). Thus, at the end of our analysis, we considered the evaluation of 75 games.

5 Findings

In this section, the findings of the studies are discussed in relation to the research questions that were defined.

5.1 RQ1: What evaluation methods have been used?

From our analysis, we identified eight main methods that are used in game evaluations, as presented below. It is observed that no new method has been identified; all the methods described in the studies are methods traditionally used in the Human-Computer Interaction area.

- **Experiment:** evaluation of game usage in a controlled environment, and conducting a statistical analysis of the results (usually involves a hypothesis test).
- **Expert review:** evaluation based on analyses by experts in ASD (e.g., therapists, teachers, and psychologists) or experts in games.
- **Focus group:** evaluation based on a group discussion facilitated by a researcher.
- **In-game evaluation:** evaluation based on game data/logs (e.g., performance data, eye tracker).
- **Interview:** evaluation based on structured and semi-structured interviews.
- **Observation:** evaluation based on observations and field notes taken during game sessions or by analyzing audio and video recordings of these sessions.
- **Pre-post test:** evaluation that collects user data before and after they play the game and then discusses and analyzes the improvements.
- **Questionnaire:** evaluations based on questionnaires that can be developed by researchers or on standard evaluation instrument(s) (e.g., Adolescent/Adult Sensory Profile questionnaire - a standard sensory profile evaluation instrument).

To classify the method(s) used for each game, we registered the method(s) explicitly mentioned by the authors, or, in case the authors did not mention a method, we classified it into one of the categories above, according to the description of the method used.

From the collected studies (i.e., 70 articles that evaluated 75 games), 5 only presented the evaluation of the algorithm

implemented in the game [Rouhi *et al.*, 2019; Frutos *et al.*, 2011; Rapela *et al.*, 2012; Dapogny *et al.*, 2018; Dantas *et al.*, 2020]. The remaining 70 games were evaluated using some method. Figure 1 shows the methods used in these evaluations. If the method was not clearly presented, the study was excluded from the visualization. For instance, in the study by Carvalho and da Cunha [2019], the authors present the parents' opinions about the game. However, they do not specify how this data was collected, so the article was not included in the graph depicted in Figure 1. Table 1 presents the classification of the games' evaluation according to the methods used. The most frequently used methods in the articles were: observation (35), followed by in-game evaluation (19), pre-post test (18), questionnaire (15), interview (11), experiment (9), expert review (7) and focus group (2).

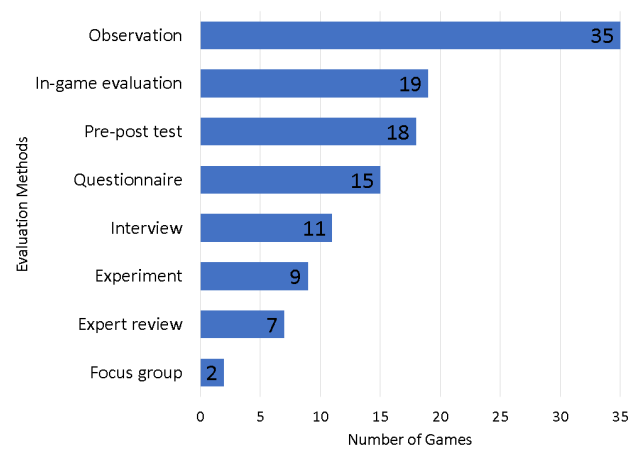


Figure 1. Number of games evaluated by each method.

5.2 RQ2: What methods have been used to evaluate each game category?

We also investigated the relationship between the methods used in evaluations and the category of the game (i.e., the skill(s) the game aims to improve). To do this, we classified the selected games into categories. The categories we used were obtained from another research we conducted, which focused on categorizing serious games for children with ASD regarding which skills they aim to develop, how their activities were operationalized, and what customization options were offered to users [de Carvalho *et al.*, 2023]. Therefore, the following categories were used:

- **Academic skills:** skills associated with school activities (e.g., reading).
- **Cognitive skills:** skills associated with information processing and knowledge application.
- **Communication skills:** skills associated with language and speech.
- **Daily living skills:** skills related to the development of self-care activities and activities required for living independently in the community (e.g., financial management).
- **Evaluation and measurement:** assessment of actions and autistic characteristics of children with ASD (e.g., sensory anomalies assessment).

Table 1. Classification of the games' evaluation according to the methods used.

| Evaluation Method | Game Reference |
|--------------------|---|
| Experiment | Chen <i>et al.</i> [2019]; Koirala <i>et al.</i> [2019]; Chan <i>et al.</i> [2016]; Li <i>et al.</i> [2018]; Rahman <i>et al.</i> [2010]; Barajas <i>et al.</i> [2017]; Hughes <i>et al.</i> [2016]; Al-Hammadi and Abdelazim [2015]; Mei <i>et al.</i> [2018] |
| Expert review | Thordarson and Vilhjálmsón [2019]; Marwecki <i>et al.</i> [2013]; Weilun <i>et al.</i> [2011]; Carvalho and da Cunha [2019]; Sousa <i>et al.</i> [2012]; Moura <i>et al.</i> [2016]; Sturm <i>et al.</i> [2016] |
| Focus group | Loiacono <i>et al.</i> [2018]; Soysa and Al Mahmud [2020] |
| In-game evaluation | Sharma <i>et al.</i> [2018](<i>Games: Ballons, HOPE, Kirana</i>), Boyd <i>et al.</i> [2018]; Ringland <i>et al.</i> [2019]; Carlier <i>et al.</i> [2019]; Spitale <i>et al.</i> [2019]; Bartoli <i>et al.</i> [2014](<i>Games: Bubble, Shape, Space</i>), Garzotto <i>et al.</i> [2014]; Wade <i>et al.</i> [2017]; Kołakowska <i>et al.</i> [2017]; Mir and Khosla [2018]; Piana <i>et al.</i> [2019]; Kurniawati <i>et al.</i> [2019]; Guerra and Furtado [2013]; Iyer <i>et al.</i> [2017]; Rodrigues <i>et al.</i> [2018] |
| Interview | Boyd <i>et al.</i> [2018]; Duval <i>et al.</i> [2018]; Ringland <i>et al.</i> [2019]; Carlier <i>et al.</i> [2019]; Gotsis <i>et al.</i> [2010]; Giusti <i>et al.</i> [2011]; Harrold <i>et al.</i> [2014]; Boyd <i>et al.</i> [2015, 2017]; Mei and Guo [2018]; Jain <i>et al.</i> [2012] |
| Observation | Loiacono <i>et al.</i> [2018]; Sharma <i>et al.</i> [2018](<i>Games: Ballons, HOPE, Kirana</i>), Boyd <i>et al.</i> [2018]; Duval <i>et al.</i> [2018]; Dragomir <i>et al.</i> [2018]; Crovari <i>et al.</i> [2019]; Aruanno <i>et al.</i> [2018]; Spitale <i>et al.</i> [2019]; Porayska-Pomsta <i>et al.</i> [2018]; Garcia-Garcia <i>et al.</i> [2019]; Buzzi <i>et al.</i> [2019]; Gotsis <i>et al.</i> [2010]; Giusti <i>et al.</i> [2011]; Harrold <i>et al.</i> [2014]; Bartoli <i>et al.</i> [2014](<i>Games: Bubble, Shape, Space</i>), Garzotto <i>et al.</i> [2014]; Boyd <i>et al.</i> [2015]; Wade <i>et al.</i> [2017]; Boyd <i>et al.</i> [2017]; Giacolini <i>et al.</i> [2015]; Marchi <i>et al.</i> [2019]; Ribeiro <i>et al.</i> [2014]; Finkelstein <i>et al.</i> [2013]; Silva and Raposo [2016]; Pistoljevic and Hulusic [2017]; Weilun <i>et al.</i> [2011]; Neto <i>et al.</i> [2013]; Rodrigues <i>et al.</i> [2018]; Carvalho and da Cunha [2019]; Gobbo <i>et al.</i> [2019]; Silva-Calpa <i>et al.</i> [2018] |
| Pre-post test | Sharma <i>et al.</i> [2018](<i>Game: Kirana</i>), Dragomir <i>et al.</i> [2018]; Carlier <i>et al.</i> [2019]; Porayska-Pomsta <i>et al.</i> [2018]; Bartoli <i>et al.</i> [2014](<i>Games: Bubble, Shape, Space</i>), Wade <i>et al.</i> [2017]; Zhang <i>et al.</i> [2018]; Boyd <i>et al.</i> [2017]; Marchi <i>et al.</i> [2019]; Piana <i>et al.</i> [2019]; Kashani-Vahid <i>et al.</i> [2018]; Hassan <i>et al.</i> [2011]; Zhao <i>et al.</i> [2018]; Uzuegbunam <i>et al.</i> [2015]; Cunha [2011]; Gobbo <i>et al.</i> [2019] |
| Questionnaire | Duval <i>et al.</i> [2018]; Aruanno <i>et al.</i> [2018]; Thordarson and Vilhjálmsón [2019]; Koirala <i>et al.</i> [2019]; Garcia-Garcia <i>et al.</i> [2019]; Buzzi <i>et al.</i> [2019]; Finkelstein <i>et al.</i> [2010]; Harrold <i>et al.</i> [2014]; Wade <i>et al.</i> [2017]; Gomez <i>et al.</i> [2018]; Golestan <i>et al.</i> [2019](<i>Games: BEESAUTI, CARAUTI</i>), Finkelstein <i>et al.</i> [2013]; Zhao <i>et al.</i> [2018]; Weilun <i>et al.</i> [2011] |

- **Motor skills:** skills related to moving oneself, or moving and interacting with objects.
- **Social and socio-emotional skills:** verbal and nonverbal skills for communication and interaction with others.
- **General:** associated with two or more skills or general purposes (i.e., not focused on a specific skill).

Table 2 presents the classification of the games according to the skills they focus on. Figure 2 presents the methods used to evaluate games in each category. The y-axis shows the skills, followed by the number of evaluations conducted focusing on this category. The x-axis contains the methods identified in the analysis (described in RQ1). It is worth noting that some studies used a combination of methods to evaluate a game. As seen in Figure 2, there is no relationship between the methods used and the skill the game focuses on. For each skill, different types of methods are used.

It is worth noting that games of *Social and socio-emotional skills* included all methods. In other categories, various evaluation methods were used, but not all. However, the cate-

gory of *Social and socio-emotional skills* has a much larger number of games than the other categories. In this case, 27 of the 70 games address this category, while 14 address the category of *General* (the 2nd most frequent).

The *questionnaire* was the only method used to evaluate all game categories. The methods of *observation* and *pre-post test* were also widely used, with the only exception being the evaluation category. However, this category has few games (only 5 of the 70 games).

5.3 RQ3: How have the methods been applied in evaluations?

To answer this question, we present two analyses. The first presents an overview of the moment in the design process the games were evaluated and in which locations, as well as how different methods were combined (5.3.1), the second presents a description of how the evaluations with each type of method was conducted (5.3.2).

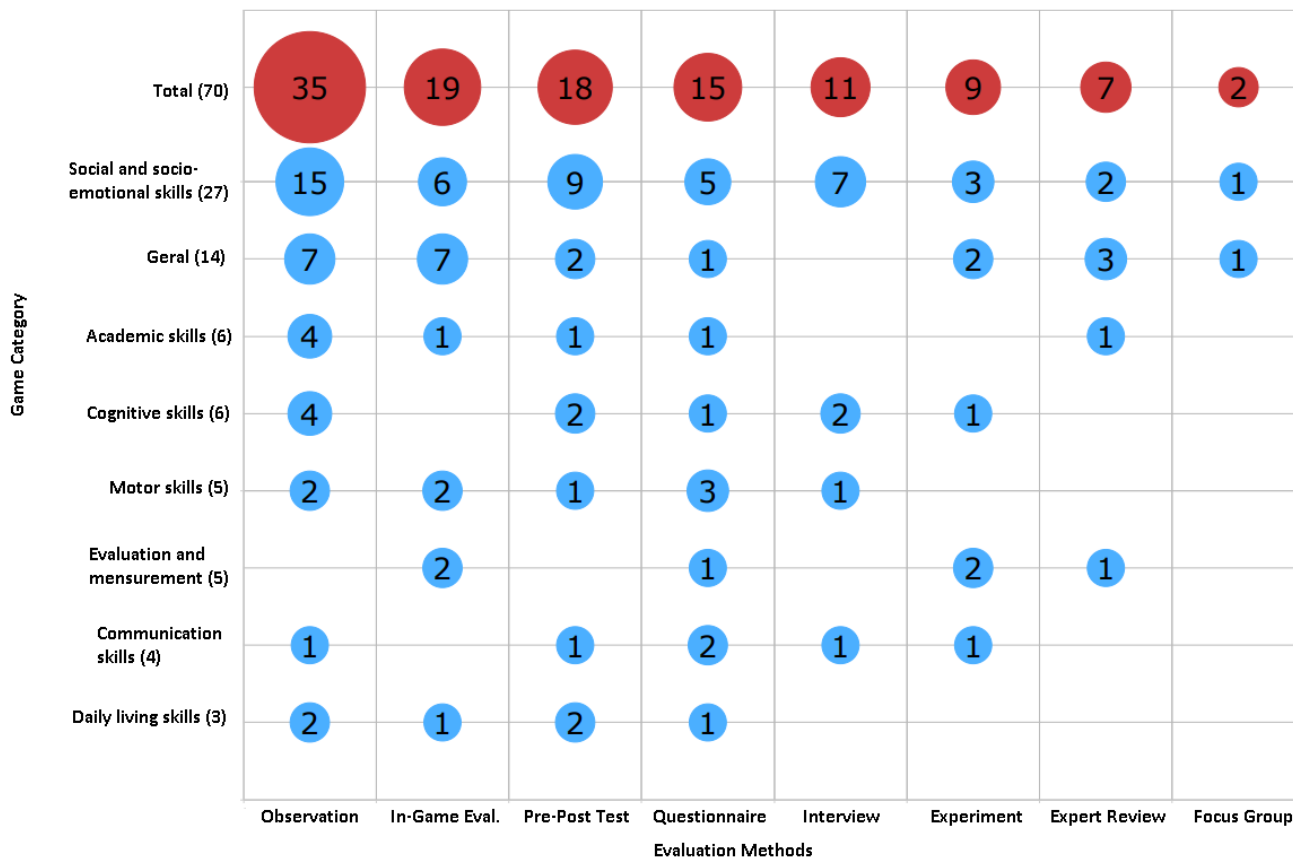


Figure 2. Evaluation methods used by game categories.

Table 2. Games’ classification according to skills they focus on.

| Skills | Game Reference |
|-----------------------------------|---|
| Academic skills | Gomez <i>et al.</i> [2018]; Kurniawati <i>et al.</i> [2019]; Pistoljevic and Hulusic [2017]; Neto <i>et al.</i> [2013]; Carvalho and da Cunha [2019]; Gobbo <i>et al.</i> [2019] |
| Cognitive skills | Dragomir <i>et al.</i> [2018]; Chen <i>et al.</i> [2019]; Buzzi <i>et al.</i> [2019]; Boyd <i>et al.</i> [2017]; Giacolini <i>et al.</i> [2015]; Mei and Guo [2018] |
| Communication skills | Duval <i>et al.</i> [2018]; Golestan <i>et al.</i> [2019],(Game: CARAUTI) Rahman <i>et al.</i> [2010]; Cunha [2011] |
| Daily living skills | Sharma <i>et al.</i> [2018](Game: Kirana), Aruanno <i>et al.</i> [2018]; Hassan <i>et al.</i> [2011] |
| Evaluation and measurement | Koirala <i>et al.</i> [2019]; Kofakowska <i>et al.</i> [2017]; Li <i>et al.</i> [2018]; Iyer <i>et al.</i> [2017]; Sturm <i>et al.</i> [2016] |
| Motor skills | Ringland <i>et al.</i> [2019]; Finkelstein <i>et al.</i> [2010]; Bartoli <i>et al.</i> [2014](Game: Bubble), Golestan <i>et al.</i> [2019](Game: BEESAUTI), Finkelstein <i>et al.</i> [2013] |
| Social and socio-emotional skills | Loiacono <i>et al.</i> [2018]; Sharma <i>et al.</i> [2018] (Game: Ballons), Boyd <i>et al.</i> [2018]; Carlier <i>et al.</i> [2019]; Porayska-Pomsta <i>et al.</i> [2018]; Thordarson and Vilhjalmsson [2019]; Garcia-Garcia <i>et al.</i> [2019]; Gotsis <i>et al.</i> [2010]; Marwecki <i>et al.</i> [2013]; Giusti <i>et al.</i> [2011]; Harrold <i>et al.</i> [2014]; Chan <i>et al.</i> [2016]; Boyd <i>et al.</i> [2015]; Wade <i>et al.</i> [2017]; Zhang <i>et al.</i> [2018]; Marchi <i>et al.</i> [2019]; Piana <i>et al.</i> [2019]; Kashani-Vahid <i>et al.</i> [2018]; Ribeiro <i>et al.</i> [2014]; Jain <i>et al.</i> [2012]; Silva and Raposo [2016]; Zhao <i>et al.</i> [2018]; Hughes <i>et al.</i> [2016]; Uzuegbunam <i>et al.</i> [2015]; Mei <i>et al.</i> [2018]; Rodrigues <i>et al.</i> [2018]; Silva-Calpa <i>et al.</i> [2018] |
| General | Sharma <i>et al.</i> [2018](Game: HOPE), Crovari <i>et al.</i> [2019]; Spitale <i>et al.</i> [2019]; Soysa and Al Mahmud [2020]; Bartoli <i>et al.</i> [2014](Games: Shape, Space), Garzotto <i>et al.</i> [2014]; Mir and Khosla [2018]; Guerra and Furtado [2013]; Barajas <i>et al.</i> [2017]; Weilun <i>et al.</i> [2011]; Al-Hammadi and Abdelazim [2015]; Sousa <i>et al.</i> [2012]; Moura <i>et al.</i> [2016] |

5.3.1 Overview of method application

The methods used in the evaluations varied according to the moment in which they were applied: some evaluated the developed game (summative evaluation), while others evaluated prototypes during the design process (formative evaluation). The evaluations also varied in relation to where they were carried out: therapeutic centers (e.g., [Loiacono *et al.*, 2018; Crovari *et al.*, 2019; Aruanno *et al.*, 2018; Spitale *et al.*, 2019]), schools (e.g., [Gomez *et al.*, 2018; Kurniawati *et al.*, 2019; Pistoljevic and Hulusic, 2017]), research laboratory (e.g., [Boyd *et al.*, 2018]) or in the participants' homes (e.g., [Ringland *et al.*, 2019; Carlier *et al.*, 2019]).

Table 3 presents the classification of the games' evaluation according to the number of methods used. Of the collected studies, 36 games were evaluated using a single evaluation method, and 34 games were evaluated using a combination of different methods, as follows: 23 games were evaluated using two methods, 10 games using three methods, and only one study combined four methods for game evaluation.

Most of the studies that used the observation method combined it with other methods (28/35). The same occurred with the in-game evaluation methods (14/19), pre-post test (13/18), questionnaire (11/15), and interview (9/11) methods. The other methods were used more as a single method in the evaluations.

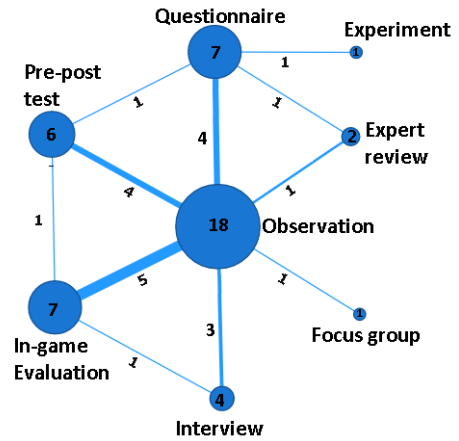
The graph in Figure 3(a) shows how the methods were combined in evaluations that used **two** methods. Each method is represented as a vertex, and its label represents the number of games evaluated using that method. The edges represent that the two methods being connected have been used together. The width and weight of the edge represent the number of games evaluated with that pair of methods. The graph in Figure 3(b) only shows the studies that used **three or four** methods in the evaluation of the games. Each method is represented as a blue round vertex, and each game evaluation is represented as a yellow square vertex. Seven different combinations of methods were used in the evaluations. The most used combination was *observation, in-game evaluation* and *pre-post test*, which was used in the evaluation of four games (Space Game, Shape Game, and Bubble Game presented in [Bartoli *et al.*, 2014], and Kirana [Sharma *et al.*, 2018]). The second most used combination was *observation, interview* and *questionnaire*, which was used in the evaluation of two games (SpokeIt [Duval *et al.*, 2018] and CopyMe [Harrold *et al.*, 2014]). Each of the other combinations of methods was used in only one study.

Section 6 details how each method was combined with others.

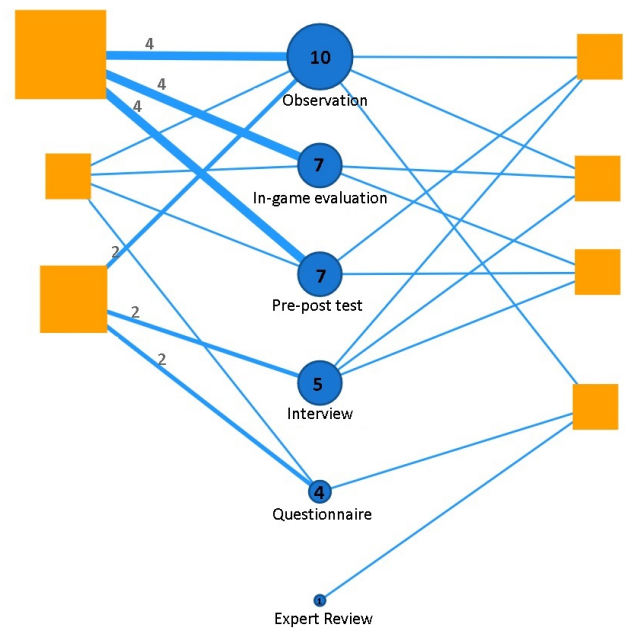
5.3.2 Description of the Evaluation Process

Next, we present an overview of the different methodologies employed in the studies. Additionally, we discuss the similarities and differences of how the methods were used.

Regarding studies that used the *observation* method, it can be pointed out that in some of them, field notes were taken, whereas in others video recordings were made, to support an in-depth analysis. Some observations were made in the context of the user and others were made in laboratories. We



(a) Games that were evaluated using a combination of two methods.



(b) Games that were evaluated using a combination of three or four methods.

Figure 3. Combination of methods in the evaluations of the games. Edges whose weight was omitted correspond to weight 1.

next describe some of these different examples of how observation was applied.

The study by Aruanno *et al.* [2018] presents HoloLearn, a wearable mixed reality application aimed at improving the ability of individuals with Neurodevelopmental Disorders (NDD) to perform typical household activities and increase their autonomy. In the exploratory study, 20 participants with NDD were divided into two groups: (i) severe-level participants (11) and (ii) moderate-level participants (9). The sessions were held at a therapeutic center, where the participants used HoloLearn with the support of their caregivers while two other professionals observed the situation and took notes. At the end of each session, the users answered a questionnaire about the device's acceptability, interaction mode usability, task complexity, virtual assistant function, friendliness, and satisfaction. In the study by Crovari *et al.* [2019], the evaluation was based on the audio and video recordings during the sessions. The researchers conducted an ex-

Table 3. Classification of the games' evaluations according to the number of methods used.

| Number of evaluation methods used | Game Reference |
|-----------------------------------|---|
| 1 evaluation method | Crovvari <i>et al.</i> [2019]; Chen <i>et al.</i> [2019]; Soysa and Al Mahmud [2020]; Finkelstein <i>et al.</i> [2010]; Marwecki <i>et al.</i> [2013]; Chan <i>et al.</i> [2016]; Kołakowska <i>et al.</i> [2017]; Gomez <i>et al.</i> [2018]; Zhang <i>et al.</i> [2018]; Giacolini <i>et al.</i> [2015]; Li <i>et al.</i> [2018]; Mir and Khosla [2018]; Mei and Guo [2018]; Kashani-Vahid <i>et al.</i> [2018]; Kurniawati <i>et al.</i> [2019]; Golestan <i>et al.</i> [2019]; Ribeiro <i>et al.</i> [2014]; Rahman <i>et al.</i> [2010]; Jain <i>et al.</i> [2012]; Hassan <i>et al.</i> [2011]; Guerra and Furtado [2013]; Silva and Raposo [2016]; Pistoljevic and Hulusic [2017]; Iyer <i>et al.</i> [2017]; Barajas <i>et al.</i> [2017]; Hughes <i>et al.</i> [2016]; Uzuegbunam <i>et al.</i> [2015]; Al-Hammadi and Abdelazim [2015]; Mei <i>et al.</i> [2018]; Neto <i>et al.</i> [2013]; Cunha [2011]; Sousa <i>et al.</i> [2012]; Moura <i>et al.</i> [2016]; Sturm <i>et al.</i> [2016]; Silva-Calpa <i>et al.</i> [2018] |
| 2 evaluation methods | Loiacono <i>et al.</i> [2018]; Sharma <i>et al.</i> [2018]; Dragomir <i>et al.</i> [2018]; Ringland <i>et al.</i> [2019]; Aruanno <i>et al.</i> [2018]; Spitale <i>et al.</i> [2019]; Porayska-Pomsta <i>et al.</i> [2018]; Thordarson and Vilhjálmsson [2019]; Koirala <i>et al.</i> [2019]; Garcia-Garcia <i>et al.</i> [2019]; Buzzi <i>et al.</i> [2019]; Gotsis <i>et al.</i> [2010]; Giusti <i>et al.</i> [2011]; Garzotto <i>et al.</i> [2014]; Boyd <i>et al.</i> [2015]; Marchi <i>et al.</i> [2019]; Piana <i>et al.</i> [2019]; Finkelstein <i>et al.</i> [2013]; Zhao <i>et al.</i> [2018]; Rodrigues <i>et al.</i> [2018]; Carvalho and da Cunha [2019]; Gobbo <i>et al.</i> [2019] |
| 3 evaluation methods | Sharma <i>et al.</i> [2018]; Boyd <i>et al.</i> [2018]; Duval <i>et al.</i> [2018]; Carlier <i>et al.</i> [2019]; Harrold <i>et al.</i> [2014]; Bartoli <i>et al.</i> [2014]; Boyd <i>et al.</i> [2017]; Weilun <i>et al.</i> [2011] |
| 4 evaluation methods | Wade <i>et al.</i> [2017] |

ploratory study to analyze how individuals diagnosed with NDD interact with SAM, an intelligent dolphin-shaped toy, and to identify requirements to improve the prototype. To stimulate autonomous exploration of SAM and active reaction to it, five individuals with NDD played and tested the game in sessions that were recorded and later analyzed. The second prototype was discussed with therapists in a workshop, and the authors observed therapeutic sessions using SAM at an institution for people with NDD. A third prototype was developed, the SAM 3D, and evaluated for the learning benefits for the users. For this purpose, a long-term empirical study (1 year) was carried out in two different therapeutic centers.

In the *in-game evaluation*, data is collected during the use of the game. Analysis of the studies showed that different types of data could be collected, both quantitative (e.g., performance in a game) and qualitative (e.g., user opinion surveys). For example, in the study by Sharma *et al.* [2018], both quantitative and qualitative data were collected to evaluate the game Balloons, which is focused on promoting social activity of joint attention, where colored balloons can be selected through gesture-based interaction. To evaluate it, data were collected regarding the total time spent by each participant in each session, the number of balloons selected with assistance, and the number of balloons selected by the children themselves. In addition, the notes of the moderator's observation were also used to validate the game. In the study [Carlier *et al.*, 2019], the game New Horizon, which aims to help reduce stress and anxiety in children with ASD, used in-game evaluation as one of the combined methods to evaluate the game. Three children were invited to play the game at their homes for two weeks. At the beginning of each game session, participants were asked to indicate their mood on a five-point Likert scale that displayed smiling faces ranging from very happy to very angry. This information was com-

bined with data collected through other methods (pre- and post-interview with parents and a standard questionnaire for parents and children) to evaluate the game's impact on the children.

Although the goal of using the *pre-post test* method is to measure the effect of the game, the approaches on what to measure and how varied greatly among the articles. Some studies applied a knowledge test at the pre and post-test stages. For example, in study [Sharma *et al.*, 2018], the authors evaluated Kirana, a game that aims to teach daily living activities of grocery shopping. The researchers combined pre-post test with observation (of the participants playing the game and in a real-life context) and in-game evaluation (game log analysis). The pre-post-test evaluation consisted of mathematics tests performed by the children at the beginning of the evaluation process and again at the end.

Other studies that employed the pre-post test method compared a control group to a treatment group. For example, in study [Bartoli *et al.*, 2014], the authors evaluated ten children with ASD regarding their initial functional profile; then, the children were randomly divided into two groups: a control group (continuation of regular treatment) and a treatment group (participation in extra sessions in which they played games); finally, the children were re-evaluated to compare the results between the two groups. Other studies compared pairs that combined children with different developmental conditions - ASD and typical development (TD). For example, in the study by Wade *et al.* [2017], the researchers organized twenty-four individuals, eighteen TD individuals and six diagnosed with ASD, into pairs as ASD-TD or TD-TD. After familiarizing the participants individually with the game in the pre-test, each pair played the game side by side sharing a computer. Players had the same game goal, individual characters and could help each other. Then, the pairs played different game modes in separate rooms but could

communicate through Skype. Finally, in the post-test stage, the pair played the same game side by side once again as in the pre-test stage. Changes in gameplay metrics, both individually and in pairs, were analyzed, as well as changes in verbal communication.

In most of the studies that used the *questionnaire* method, the researchers created their questionnaire form to assess the usability and user experience of the participants (12/15). Half of them (6 articles) reported using the Likert scale. Only three studies used standardized questionnaires with different focuses, namely: Adolescent/Adult Sensory Profile (AASP) [Koirala *et al.*, 2019], System Usability Scale (SUS) [Garcia-Garcia *et al.*, 2019], and Social Communication Questionnaire (SCQ) and Social Responsiveness Survey, Second Edition (SRS-2) [Wade *et al.*, 2017]. For example, the study by Koirala *et al.* [2019] used a standard questionnaire combined with an experiment to validate the feasibility of the Sensory Assessment VR System (SAVR), a game for assessing sensory abnormalities in children with ASD. Before playing the game, all participants (six children with ASD and six typically developing children) filled out an Adolescent/Adult Sensory Profile (AASP) questionnaire, a standard sensory profile assessment tool. Then, the children interacted with the game. The researchers analyzed if the results from the SAVR system were or were not correlated with the results from the AASP questionnaire, which would indicate alignment between the measure constructed within the system and the sensory profile of the participants evaluated by a standard psychological instrument.

The *interviews* were conducted in various ways and involved different participants in each study, such as children, parents, therapists, teachers, and even game development students. For example, the study by Mei and Guo [2018] conducted a preliminary evaluation of an adaptive virtual environment therapy system for children with autism spectrum attention deficit. To collect feedback on the system, two ASD therapists and six game development students aged 19 to 22 were interviewed. The therapists evaluated the game's potential to be used in therapy, and the students evaluated attention detection.

The studies that applied the *experiment* method conducted evaluations that collected metrics as participants played the game. In some studies, one or more hypotheses were raised, and the experiment aimed to prove or reject each of them (e.g., [Chen *et al.*, 2019; Mei *et al.*, 2018]). Other studies did not present hypotheses but collected data, performed a statistical analysis of this data, and then discussed the results (e.g., [Li *et al.*, 2018; Rahman *et al.*, 2010]). To illustrate, in the study [Chen *et al.*, 2019], the researchers conducted an experiment to evaluate the *Guided Play Blocks*, a game to improve symbolic play skills in children with autism spectrum disorder. The aim was to verify the following hypotheses: (H1) a child with restrictive and repetitive behaviors in a physical activity can also exhibit similar patterns in a digital replica of the activity, and (H2) interventions carried out with digital activity can impact the child's behavior in the physical world. Six children with autism spectrum disorder participated in an experiment that involved playing the game in the free and guided modes. Quantitative and qualitative data were collected for analysis. The game mode was

the independent variable tested (i.e., free game mode versus guided game mode), and the dependent variables considered were: i) the percentage of representational constructions (i.e., a construction made with the blocks that resembles a real-life object); ii) the number of representational constructions; iii) the total number of symbolic categories of the constructions (i.e., a category consists of constructions with the same symbolic meaning, e.g., animals, letters); and iv) the compliance with the guidelines (i.e., captures how well a child follows the system's guidance).

The works that conducted *expert reviews* typically involve specialists in ASD (therapists, teachers, and psychologists) [Thordarson and Vilhjálmsón, 2019; Weilun *et al.*, 2011; Carvalho and da Cunha, 2019; Moura *et al.*, 2016; Sturm *et al.*, 2016] or HCI [Sousa *et al.*, 2012], or both [Marwecki *et al.*, 2013]. For example, to identify usability problems in the TEO [Moura *et al.*, 2016] - a suite of interactive games aimed at helping the learning of various fundamental concepts about ASD treatment - five therapists and psychologists were invited to conduct an evaluation. The specialists tested the TEO and answered a questionnaire about the game. In the study by Sousa *et al.* [2012], the WorldTour game was developed to support the learning process of children with TEA. HCI specialists evaluated the game's usability and communicability using: Heuristic Evaluation, the Semiotic Inspection Method (SIM), and the Communicability Evaluation Method (CEM).

The studies that used *focus groups* carried out the evaluations with specialists in TEA, such as teachers and therapists [Loiacono *et al.*, 2018; Soysa and Al Mahmud, 2020]. For example, in the study [Soysa and Al Mahmud, 2020], five special education teachers and 20 children with ASD analyzed the usability of a tangible interface-based game prototype. The children formed pairs with the teachers, who chose the activities to be performed, according to the needs of each child, to play the game. At the end of the sessions, a focus group was held with the professionals to discuss the difficulties faced during the gaming time and identify improvements in the game.

5.4 RQ4: What is the sample size and profile of the participants in the evaluations?

The studies conducted their evaluations with different stakeholders (e.g., therapists, psychologists, and teachers) and not just with the intended users, that is, individuals with ASD or with NDD. In this section, we have analyzed the participants of the evaluations and organized our results in two sub-sections - the first regarding the participants who were end-users, and the other, regarding evaluations with stakeholders.

5.4.1 Characterizing End-User Participants

For the evaluations that included users in their evaluation, we performed an analysis to characterize the end-user evaluations according to the number of participants involved, their profile in terms of their special needs, and their age. Articles that did not include information about the participants when presenting the evaluation were not considered. A total of 67

evaluations were described, with two different evaluations conducted for one of the games [Marchi *et al.*, 2019]; one evaluation was not considered [Sturm *et al.*, 2016] because it did not describe the number and age range of the participants. Table 4 presents the references of the games' evaluations according to the end-user participants' profiles in the game evaluation.

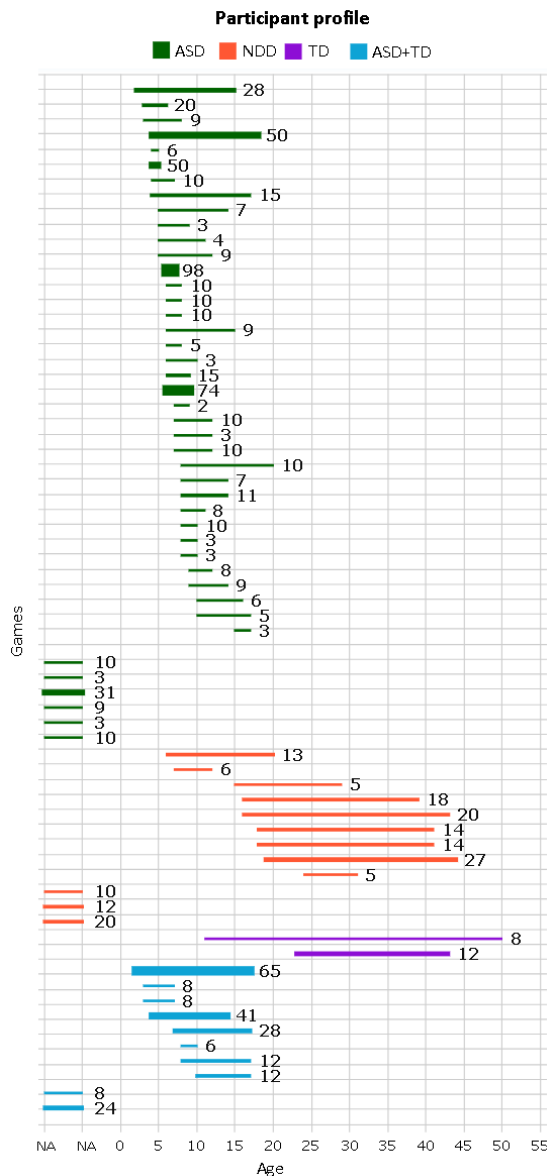


Figure 4. Characterization of the end-user participants in the evaluations. Columns NA: the games that did not specify the age range.

Figure 4 presents a visualization that shows, for each evaluated game, the number, profile, and age range of the evaluation participants. The vast majority of the studies performed evaluations with participants representative of the target audience with ages ranging from 2 to 20 years, but concentrated in the range between 5 and 15 years. Although our SLR focused on games for children with ASD, we can see in Figure 4 that evaluations of nine games also included adults⁵. Some of the games described included children in their target user group but were not intended **only** for chil-

⁵We consider adults any participant over twenty-one years old.

dren (e.g., [Crovari *et al.*, 2019], [Aruanno *et al.*, 2018], [Spitale *et al.*, 2019], and 2/3 games - Balloons and HOPE - in [Sharma *et al.*, 2018]). Thus, in these cases, it made sense to also include adults in their evaluation. However, four studies [Duval *et al.*, 2018; Thordarson and Vilhjálmsón, 2019; Finkelstein *et al.*, 2010; Sharma *et al.*, 2018] presented games specifically aimed at children but performed the evaluations with adults. The authors of these articles [Sharma *et al.*, 2018; Thordarson and Vilhjálmsón, 2019; Finkelstein *et al.*, 2010] did not give any reasons for this choice. In contrast, article [Duval *et al.*, 2018] argued that although the game Spokelt was made for children with developmental and speech disabilities, it was difficult to access these individuals; thus, their solution was to test with adults with this profile.

The sample size of participants ranged from 2 to 98, with an average sample size of 15, with a standard deviation of 17. The majority of evaluations were carried out with individuals with ASD, but some studies also included individuals with NDD and neurotypical individuals. The majority of evaluations with individuals with NDD included at least one person with ASD (7/12), but the remaining five works did not specify the developmental disorders of the participants. Moreover, 10 studies carried out evaluations that encompassed not only individuals with ASD, but also individuals with TD [Porayska-Pomsta *et al.*, 2018; Koirala *et al.*, 2019; Harrold *et al.*, 2014; Wade *et al.*, 2017; Zhang *et al.*, 2018; Li *et al.*, 2018; Golestan *et al.*, 2019; Zhao *et al.*, 2018; Carvalho and da Cunha, 2019; Sturm *et al.*, 2016].

The studies by Porayska-Pomsta *et al.* [2018], Koirala *et al.* [2019], Li *et al.* [2018] and Sturm *et al.* [2016] explained that the reason for choosing both profiles was to identify and compare differences between them (e.g., behavioral differences, sensory processing patterns). On the other hand, the works by Wade *et al.* [2017], Zhang *et al.* [2018], and Zhao *et al.* [2018] developed and evaluated games that focused on social interaction and communication and organized the participants into ASD-TD dyads, and also TD-TD [Wade *et al.*, 2017; Zhang *et al.*, 2018]. The reason pointed out by the researchers for not grouping the participants in ASD-ASD pairs was that the goal of the intervention included improving relationships between individuals with ASD and their neurotypical peers, since it is more common for people with ASD to interact with neurotypical individuals in their daily lives. In relation to the studies that organized the participants into TD-TD pairs, it was described that this allowed researchers to identify differences between the interactions of neurotypical and ASD individuals, and also allowed verifying the applicability of the games for individuals with and without ASD.

The remaining studies that included the two groups did not explain the reasons for this evaluation configuration [Harrold *et al.*, 2014; Golestan *et al.*, 2019; Carvalho and da Cunha, 2019]. In two studies, the evaluations were carried out only with neurotypical individuals. Neither of them explained why the games were not tested with the respective target audience, but indicated this objective as future work [Thordarson and Vilhjálmsón, 2019; Finkelstein *et al.*, 2010].

The visualization depicted in Figure 5 indicates the total number of studies that used a given evaluation method that in-

Table 4. Classification of the games’ evaluations, according to the profile of the evaluation end-user participant’s development (ASD - Autism Spectrum Disorder; NDD - Neurodevelopmental Disorders; TD - Typical Development) included in the evaluations.

| Participant Profile | Game Reference |
|---------------------|--|
| ASD | Boyd <i>et al.</i> [2018]; Dragomir <i>et al.</i> [2018]; Ringland <i>et al.</i> [2019]; Carlier <i>et al.</i> [2019]; Chen <i>et al.</i> [2019]; Garcia-Garcia <i>et al.</i> [2019]; Buzzi <i>et al.</i> [2019]; Soysa and Al Mahmud [2020]; Gotsis <i>et al.</i> [2010]; Giusti <i>et al.</i> [2011]; Bartoli <i>et al.</i> [2014](<i>Games: Bubble, Shape, Space</i>), Garzotto <i>et al.</i> [2014]; Boyd <i>et al.</i> [2015]; Kołakowska <i>et al.</i> [2017]; Gomez <i>et al.</i> [2018]; Giacolini <i>et al.</i> [2015]; Marchi <i>et al.</i> [2019]; Mir and Khosla [2018]; Piana <i>et al.</i> [2019]; Ribeiro <i>et al.</i> [2014]; Rahman <i>et al.</i> [2010]; Finkelstein <i>et al.</i> [2013]; Jain <i>et al.</i> [2012]; Hassan <i>et al.</i> [2011]; Guerra and Furtado [2013]; Silva and Raposo [2016]; Pistoljevic and Hulusic [2017]; Iyer <i>et al.</i> [2017]; Barajas <i>et al.</i> [2017]; Hughes <i>et al.</i> [2016]; Uzuegbunam <i>et al.</i> [2015]; Weilun <i>et al.</i> [2011]; Al-Hammadi and Abdelazim [2015]; Mei <i>et al.</i> [2018]; Neto <i>et al.</i> [2013]; Cunha [2011]; Rodrigues <i>et al.</i> [2018]; Gobbo <i>et al.</i> [2019]; Sousa <i>et al.</i> [2012]; Silva-Calpa <i>et al.</i> [2018] |
| NDD | Loiacono <i>et al.</i> [2018]; Sharma <i>et al.</i> [2018](<i>Games: Ballons, HOPE, Kirana</i>), Duval <i>et al.</i> [2018]; Crovari <i>et al.</i> [2019]; Aruanno <i>et al.</i> [2018]; Spitale <i>et al.</i> [2019]; Chan <i>et al.</i> [2016]; Boyd <i>et al.</i> [2017]; Kashani-Vahid <i>et al.</i> [2018]; Kurniawati <i>et al.</i> [2019] |
| ASD+TD | Porayska-Pomsta <i>et al.</i> [2018]; Koirala <i>et al.</i> [2019]; Harrold <i>et al.</i> [2014]; Wade <i>et al.</i> [2017]; Zhang <i>et al.</i> [2018]; Li <i>et al.</i> [2018]; Golestan <i>et al.</i> [2019](<i>Games: BEESAUTI, CARAUTI</i>), Zhao <i>et al.</i> [2018]; Carvalho and da Cunha [2019]; Sturm <i>et al.</i> [2016] |
| TD | Thordarson and Vilhjálmsón [2019]; Finkelstein <i>et al.</i> [2010] |

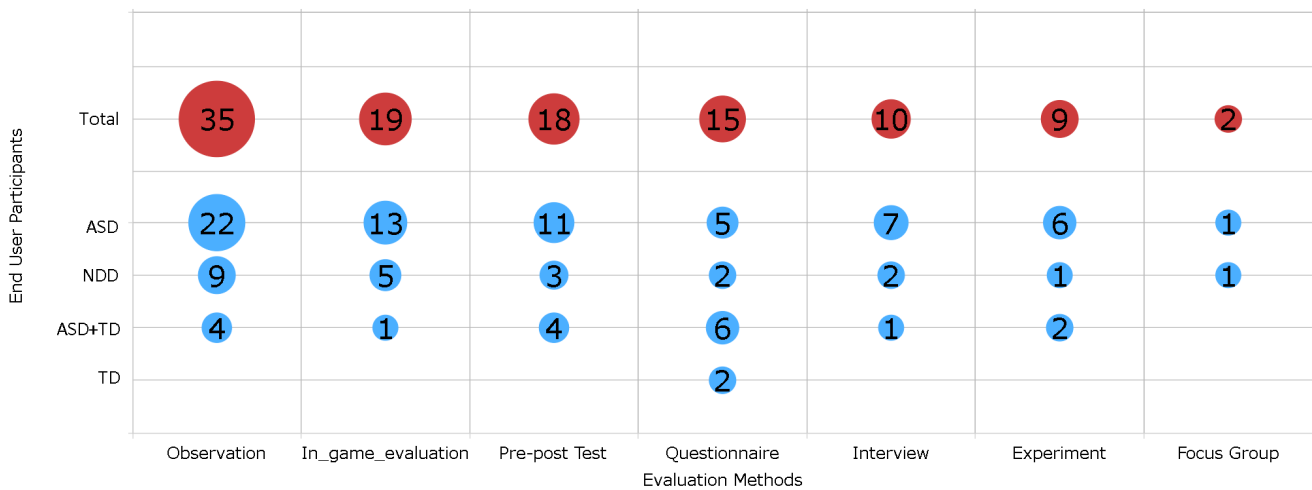


Figure 5. Characterization of evaluations performed with each method and each user profile.

cluded end-users and the profile of the participants involved. For instance, Figure 5 shows that 35 studies applied the *observation* method, and of these applications: 22 studies applied this method with participants with ASD; 9 studies conducted observation with participants with NDD, and 4 studies with participants with ASD as well as TD. It is important to remember that some studies used a combination of methods to evaluate a serious game, so they will be counted for each method applied.

We can observe that most of the methods involved participants with ASD. The *questionnaire* method was the only method that included applications with only TD participants. As previously reported, these two studies did not describe why the games were not tested with the game’s target audience [Thordarson and Vilhjálmsón, 2019; Finkelstein *et al.*, 2010]. The *questionnaire* was the most used method with users with ASD+TD. In these cases, the goal was to iden-

tify and compare differences between profiles [Koirala *et al.*, 2019] or evaluate games focused on children’s interaction and social communication [Harrold *et al.*, 2014; Golestan *et al.*, 2019]. Two studies did not describe the motivation to use participants with ASD+TD [Wade *et al.*, 2017; Zhao *et al.*, 2018].

We can also observe that the techniques that involve less intervention by the evaluator during the evaluation, for example, *observation* and *in-game evaluation*, had a more significant number of studies that explored evaluations with ASD and NDD participants. In contrast, methods such as *questionnaire* and *interview*, which demand greater interaction between the evaluator and the participants were less used with ASD and NDD participants. A possible reason for this, could be the challenges in collecting data through direct communication with the target audience.

Regarding the focus groups, their intent was to collect data

on end-users experience, albeit indirectly. For instance, in the study by Soysa and Al Mahmud [2020], 20 children with ASD used the serious game POMA with the supervision of special education teachers. At the end of the sessions, a focus group was conducted with the teachers to discuss the problems faced by the children when playing the game and to identify possible improvements to mitigate such problems. Likewise, in the article by Loiacono *et al.* [2018], 10 children with NDD played a memory-like virtual reality game to enhance social skills, with the supervision of their therapists. At the end of the study, two focus groups were conducted with NDD specialists to understand the children and therapists' needs and identify the main characteristics and parameters of the virtual reality experience that were critical during its use.

5.4.2 Characterizing Stakeholder Participants

We also performed an analysis of the stakeholders involved in game evaluations from whom data was collected⁶. In this case, the data was not collected directly from the end-users, but rather from other people who had an interest in the game, such as parents, therapists, psychologists, and teachers.

For example, in the studies, the stakeholders had the following actions: i) in the evaluations with the *questionnaire* method, the stakeholders answered the questionnaire; ii) in evaluations using the *interview* method, stakeholders were interviewed; iii) in the evaluations using the *focus group* method, the participants participated in the focus group discussion; iv) in the evaluations with the *pre-post test* method, the stakeholders played the game together with the participants or provided information about the participants during the data collection performed before and after the participant played the game, and v) in the evaluations with the *expert review* method, stakeholders analyzed the game.

Figure 6 presents a visualization that shows, for each evaluation method, the number and profile of stakeholders involved in the evaluation of each game. The following stakeholders participated in the evaluations: caregiver, educational professional (it includes teacher, educator, special education teacher, and educational adviser), game development students, interact expert, lecturer in game analysis, parents, practitioner⁷, professional designer of games, psychologist, and therapist (it includes occupational therapist, autism therapist, and NDD specialist). We included a *psychologist+therapist* profile in the visualization since the study presented by Moura *et al.* [2016] only described that five psychologists and therapists participated in the evaluation but did not detail the exact number of participants with each of these profiles.

The number of stakeholders involved in the studies ranged from 1 to 9 individuals. These studies used 5 of the 8 identified evaluation methods: *expert review*, *focus group*, *interview*, *questionnaire*, and *pre-post test*. Studies with the *expert review* method used a greater variety of stakeholder profiles, which may have occurred due to the nature of the technique. In evaluations with experts, there is a tendency

for the study to include different profiles of domain experts. The interview and questionnaire involved a more significant number of parents. These techniques require direct interaction between the evaluator and the participant. When it is impossible to apply these methods directly with the children, the parents can be the children's representatives. The challenges in including the children directly, could be due to the vulnerable nature of ASD or NDD children, or even, in some cases, their difficulties to communicate. Parents and education professionals were the stakeholder profiles that were included more often in the evaluations, probably because they are the stakeholders with the most significant interaction with the games' target audience.

5.5 RQ5: What quality properties have been evaluated in the studies?

This research question aimed to investigate which quality properties the researchers considered in their evaluations of serious games. Next, we present the seven quality properties we identified from our analysis and their description:

- **Communicability:** evaluations that analyze the serious game's communicability (i.e., the game's ability to effectively and efficiently convey to the user the intentions and interaction principles that guided its design) [de Souza, 2005].
- **Engagement:** evaluations that investigate the serious game's ability to involve the user.
- **Feasibility:** evaluations that investigate whether the serious game is suitable for use in therapy, both in terms of the serious game's proposal and in terms of acceptance of use due to some physical device that the serious game requires or due to the type of interaction (e.g., immersion in virtual or augmented reality).
- **Impact:** evaluations that investigate the effects of the serious game on users in relation to the skill the serious game aims to enhance.
- **Metrics effectiveness:** encompasses both studies that propose metrics and evaluate them based on serious games (i.e., verify whether the proposed metrics can achieve the goal of measuring some autistic characteristic of users) and studies that analyze the impact of a specific component of the serious game.
- **Usability:** evaluations that investigate the serious game's solution proposal (i.e., evaluation of the game's functionalities), the ease of use of the serious game, or the user's performance in the task proposed in the serious game. It is worth noting that evaluations that focused on analyzing the user's performance in a task and superficially investigated user satisfaction, for example, through a questionnaire, were classified as usability evaluations.
- **User experience:** evaluations that investigate the user's behavior and emotions when using the serious game. This category also included evaluations that more deeply investigated user satisfaction.

To classify the quality property(ies) evaluated in each game, we registered the property explicitly mentioned by

⁶In some cases, stakeholders helped with the use of the game during an evaluation, but their perspectives or views were not collected.

⁷The study did not explain the background of the practitioner.

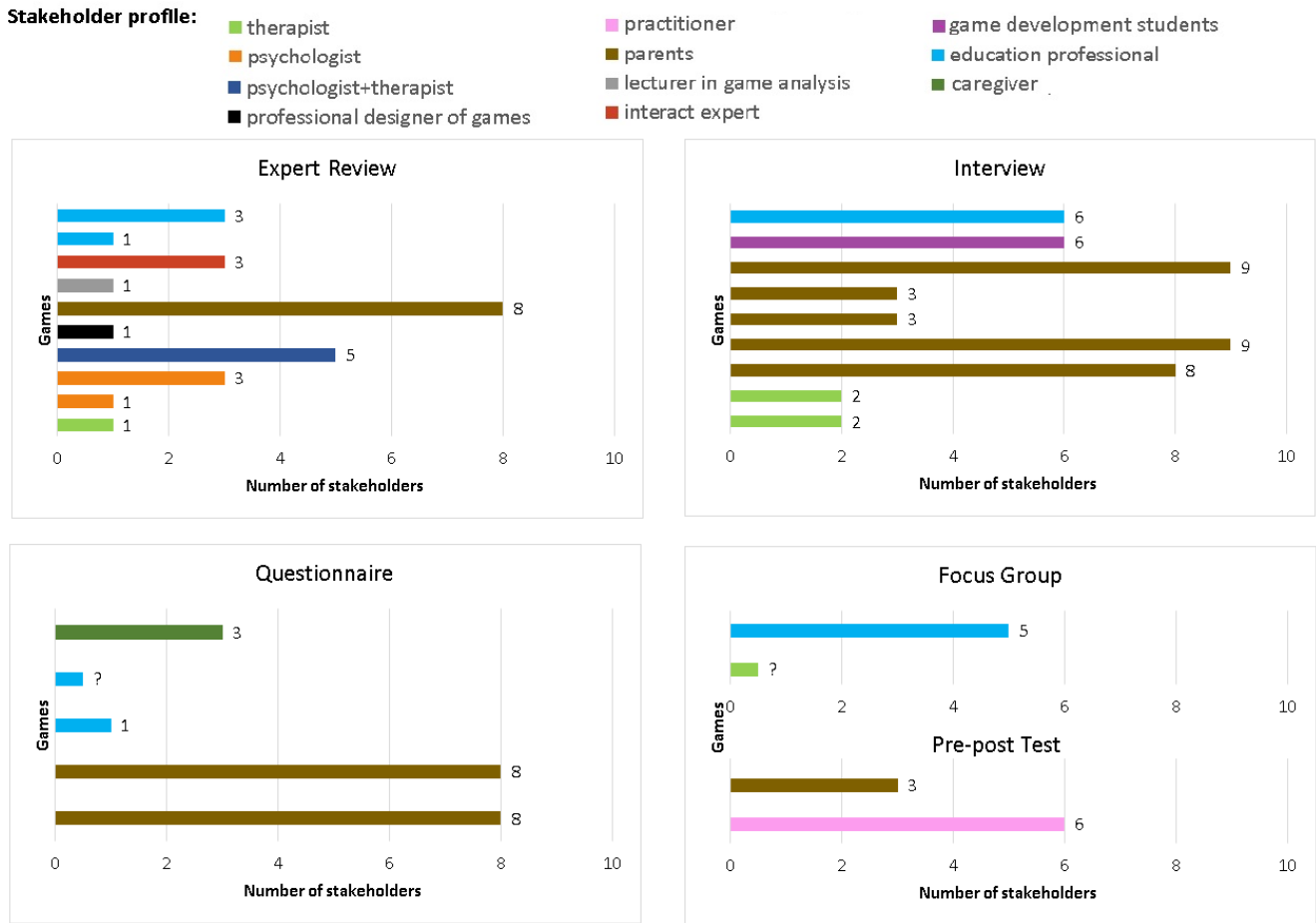


Figure 6. Characterization of the stakeholders involved in the evaluations. Marker ?: studies that did not specify the number of stakeholders involved in the evaluation.

the researchers. If the researchers did not directly mention this information, we analyzed the property being considered based on the description of the evaluation’s objective, the process conducted in the study, and the results obtained. Table 5 presents the classification of the games’ evaluation according to the quality property considered.

Figure 7 presents the quality properties evaluated in the games. The most frequently investigated properties considered in the evaluations were: impact (29), usability (22), feasibility (14), metric effectiveness (5), user experience (5), engagement (4), and communicability (1).

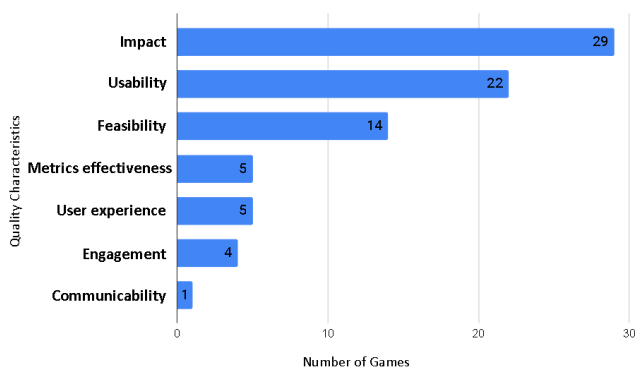


Figure 7. Quality properties evaluated in the games.

The evaluations that investigated the *impact* of the games involved both quantitative analyses that measured metrics related to the game’s objective, as well as qualitative ones. About the evaluations that analyzed *metrics effectiveness*, most (4/5) are of the evaluation category [Koirala *et al.*, 2019; Kolakowska *et al.*, 2017; Li *et al.*, 2018; Iyer *et al.*, 2017], i.e., focusing on metrics generated by the game that evaluate autistic characteristics of children. Only one article in this category analyzed the impact of a game component, namely the impact of having a custom or non-custom virtual character on engagement [Mei *et al.*, 2018].

6 Characterizing the Use of the Evaluation Methods

While in section 5, through a discussion of research questions, we present an overview of evaluation methods showing the different relationships between them. In this section, we present each evaluation method in depth. Thus, we present for each method: i) the categories of games that were evaluated; ii) the methods that were combined with it; iii) the profile of the participants and stakeholders involved in the evaluations; iv) the sample size (minimum, maximum, and average size) and age range of the participants; and v) the evaluated quality properties. It is worth noting that most

Table 5. Classification of the games' evaluations according to the quality property considered.

| Quality properties | Game Reference |
|-----------------------|--|
| Communicability | [Sousa <i>et al.</i> , 2012] |
| Engagement | Ringland <i>et al.</i> [2019]; Crovari <i>et al.</i> [2019]; Gotsis <i>et al.</i> [2010]; Finkelstein <i>et al.</i> [2013] |
| Feasibility | Boyd <i>et al.</i> [2018]; Aruanno <i>et al.</i> [2018]; Thordarson and Vilhjálmsón [2019]; Marwecki <i>et al.</i> [2013]; Giusti <i>et al.</i> [2011]; Chan <i>et al.</i> [2016]; Gomez <i>et al.</i> [2018]; Zhang <i>et al.</i> [2018]; Mei and Guo [2018]; Golestan <i>et al.</i> [2019]; Ribeiro <i>et al.</i> [2014]; Weilun <i>et al.</i> [2011]; Silva-Calpa <i>et al.</i> [2018] |
| Impact | Sharma <i>et al.</i> [2018]; Boyd <i>et al.</i> [2018]; Dragomir <i>et al.</i> [2018]; Carlier <i>et al.</i> [2019]; Spitale <i>et al.</i> [2019]; Porayska-Pomsta <i>et al.</i> [2018]; Chen <i>et al.</i> [2019]; Harrold <i>et al.</i> [2014]; Bartoli <i>et al.</i> [2014]; Garzotto <i>et al.</i> [2014]; Boyd <i>et al.</i> [2015]; Wade <i>et al.</i> [2017]; Boyd <i>et al.</i> [2017]; Marchi <i>et al.</i> [2019]; Mir and Khosla [2018]; Piana <i>et al.</i> [2019]; Kashani-Vahid <i>et al.</i> [2018]; Rahman <i>et al.</i> [2010]; Hassan <i>et al.</i> [2011]; Pistoljevic and Hulusic [2017]; Zhao <i>et al.</i> [2018]; Barajas <i>et al.</i> [2017]; Hughes <i>et al.</i> [2016]; Uzuegbunam <i>et al.</i> [2015]; Al-Hammadi and Abdelazim [2015]; Cunha [2011]; Gobbo <i>et al.</i> [2019] |
| Metrics effectiveness | Koirala <i>et al.</i> [2019]; Kofakowska <i>et al.</i> [2017]; Li <i>et al.</i> [2018]; Iyer <i>et al.</i> [2017]; Mei <i>et al.</i> [2018] |
| Usability | Loiacono <i>et al.</i> [2018]; Sharma <i>et al.</i> [2018]; Duval <i>et al.</i> [2018]; Aruanno <i>et al.</i> [2018]; Thordarson and Vilhjálmsón [2019]; Garcia-Garcia <i>et al.</i> [2019]; Buzzi <i>et al.</i> [2019]; Soysa and Al Mahmud [2020]; Finkelstein <i>et al.</i> [2010]; Gotsis <i>et al.</i> [2010]; Giacolini <i>et al.</i> [2015]; Kurniawati <i>et al.</i> [2019]; Guerra and Furtado [2013]; Pistoljevic and Hulusic [2017]; Weilun <i>et al.</i> [2011]; Rodrigues <i>et al.</i> [2018]; Carvalho and da Cunha [2019]; Gobbo <i>et al.</i> [2019]; Sousa <i>et al.</i> [2012]; Moura <i>et al.</i> [2016]; Sturm <i>et al.</i> [2016] |
| User experience | Wade <i>et al.</i> [2017]; Jain <i>et al.</i> [2012]; Silva and Raposo [2016]; Zhao <i>et al.</i> [2018]; Neto <i>et al.</i> [2013] |

studies that used a combination of methods do not specify the quality properties evaluated through each method. Instead, the studies usually describe the quality properties that were investigated in the work as a whole, using the set of evaluation methods. Next, we present the evaluation methods organized from the most frequently used in the studies, to the least used.

6.1 Observation

The *observation* method was used to evaluate games from 7 out of the 8 identified game categories. It was only not used to evaluate games from the *evaluation and measurement* category. Among the 35 studies that applied the *observation* method, 28 used this method in combination with other evaluation methods, and 7 studies applied *observation* in an isolated manner.

In cases where the method was used in conjunction with other evaluation approaches, it was combined with one, two, or three other distinct methods. In 18 studies, it was used in conjunction with only one other method, with 5 studies combining it with *in-game evaluation*; 4 studies with *pre-post test*; 4 studies with *questionnaire*; 3 studies with *interview*; 1 study with *focus group*; and 1 study with *expert review*. Another 9 studies combined the application of the *observation* method with two methods, with 4 studies combining it with *in-game evaluation* and *pre-post test*; 2 studies combin-

ing it with *interview* and *questionnaire*; 1 study combining it with *pre-post test* and *interview*; and 1 study combining it with *in-game evaluation* and *interview*. In a single study, the *observation* method was combined with three other methods, namely *in-game evaluation*, *pre-post test*, and *questionnaire*.

In studies where the *observation* method was used, both in isolation and in combination with other approaches, the sample size varied from 3 to 41 participants, with an average of 10.7 participants (standard deviation of 17.4). The ages of the participants ranged from 2 to 44 years. In most of these studies, the participants were individuals with ASD (22 studies), followed by NDD (9 studies) and ASD+TD (4 studies). No stakeholders were involved in the evaluations with the *observation* method.

With regard to the evaluated quality properties, studies that used the *observation* method in an isolated manner assessed the following set of quality properties: user experience (2), feasibility (2), engagement (1), usability (1), and impact (1). In turn, studies in which the *observation* method was combined with other evaluation approaches evaluated the same five quality properties: impact (15), usability (12), feasibility (4), engagement (2), and user experience (1).

6.2 In-game Evaluation

The *In-game evaluation* method was used to evaluate games in the following categories: general (7), social and socio-

emotional skills (6), motor skills (2), evaluation and measurement (2), academic skills, and daily living skills (1). Most studies collected data on the participant's performance when using the game. But it's important to note that the data collected highly depends on the game. For example, the game Kirana [Sharma *et al.*, 2018] - a game that aims to teach activities of daily living in grocery shopping - collected task times, items bought, and monetary transaction details. The vrSocial game [Boyd *et al.*, 2018] - an immersive virtual reality game aimed at improving the social communication skills of children with ASD - collected each user's distance from the avatar, volume, and duration of talking.

Among the 19 studies that applied the *In-game evaluation* method, the majority (14/19) used this method in combination with other methods. In these cases, it was combined with one, two, or three other distinct methods. In 7 studies, the *In-game evaluation* method was used with another method, with 5 studies combining it with *observation*; 1 study with *interview*; and 1 study with *pre-post test*. Other studies combined the use of the *In-game evaluation* method with two other methods, with 4 studies combining it with *observation* and *pre-post test*; 1 study combining it with *observation* and *interview*; and 1 study combining it with *pre-post test* and *interview*. In a single study, the *In-game evaluation* method was combined with three other methods, namely *observation*, *pre-post test*, and *questionnaire*.

In studies where the *In-game evaluation* method was used (either alone or in combination with other methods), the average sample size was 16.6 participants, with a standard deviation of 17.9. The sample size ranged from 3 to 50 participants, with ages ranging from 4 to 44 years. In most of these studies, the participants were individuals with ASD (13 studies), followed by NDD (5 studies) and ASD+TD (1 study). No stakeholders were involved in the evaluations using the *In-game evaluation* method.

Regarding the evaluated quality properties, the 5 studies that used the *In-game evaluation* method as the only method evaluated the following set of quality properties: metrics effectiveness (2 studies), usability (2 studies), and impact (1 study). In studies where the *In-game evaluation* method was combined with other evaluation approaches, the following quality properties were evaluated: impact (10 studies), usability (3 studies), engagement (1 study), feasibility (1 study), and user experience (1). It is noteworthy that two studies evaluated two quality properties.

6.3 Pre-post test

The *pre-post test* method was used to evaluate games from 7 of the 8 identified categories of games. The only category from which it was not used to evaluate games was the *evaluation and measurement* category. The *pre-post test* method was used in 18 studies, with 13 studies using it in conjunction with other evaluation approaches and 5 studies using it in isolation.

In 6 studies, the *pre-post test* method was used in conjunction with only one other method: *observation* (4 studies); *in-game evaluation* (1 study); and *questionnaire* (1 study). Other 6 studies combined the *pre-post test* method with two other methods, with 4 studies combining it with *observa-*

tion and *in-game evaluation*; 1 study combining it with *in-game evaluation* and *interview*; and 1 study combining it with *observation* and *interview*. One study combined the *pre-post test* method with three methods, namely *observation*, *in-game evaluation*, and *questionnaire*.

In the studies that used the *pre-post test* method, the average sample size was 13.2 participants, with a standard deviation of 17.8. The participant sample size ranged from 3 to 28 participants. The ages of the participants ranged from 2 to 39 years old. The participants had the following profiles: ASD (11 studies), NDD (3 studies), and ASD+TD (4 studies). Two studies [Dragomir *et al.*, 2018; Carlier *et al.*, 2019] also involved stakeholders in the evaluations with the *pre-post test* method. In the study by Dragomir *et al.* [2018], a game was presented and evaluated to help autistic children engage in solitary pretend play. The evaluation involved the participation of 6 practitioners and 7 children with ASD. The children participated in 5 play sessions over 5 weeks. In the first and last sessions, the children played with a pre-defined set of toys alone and then with the practitioner. In the intermediate sessions, each child and a practitioner played the game.

Regarding the evaluated quality properties, studies that used the *pre-post test* method in isolation evaluated the following quality properties: impact (4 studies) and usability (1 study). Studies that combined the *pre-post test* method with other methods evaluated the following set of quality properties: impact (13 studies), user experience (2 studies), and usability (1 study). Two of these studies evaluated more than one quality property.

6.4 Questionnaire

The *questionnaire* method was the only method used to evaluate games in all identified categories. Among the 15 studies that applied the *questionnaire* method, 11 used this method in combination with other evaluation methods, and 4 studies applied the *questionnaire* method in isolation.

In 7 studies, it was used in combination with only one other method, with 4 studies combining it with *observation*; 1 study with *pre-post test*; 1 study with *experiment*; and 1 study with *expert review*. Another 3 studies combined the application of the *questionnaire* method with two methods, with 2 studies combining it with *observation* and *interview*; and 1 study combining it with *observation* and *expert review*. Only one study combined the *questionnaire* method with three other methods, namely *observation*, *in-game evaluation*, and *pre-post test*.

In the 15 studies where the *questionnaire* method was used, the sample size ranged from 3 to 24 participants, with an average size of 9.9 participants (standard deviation of 15.4). The participants were aged between 3 and 50 years. Regarding the profile of the participants, 6 studies involved individuals with ASD and TD, 5 studies involved only individuals with ASD, 2 studies involved individuals with NDD, and 2 studies involved only participants with TD. Some of the studies that used the *questionnaire* method also involved stakeholders in the evaluations. In these studies, the questionnaires were filled out only by stakeholders or were filled out by both stakeholders and end users. In three evaluations, the

questionnaire was filled out only by stakeholders.

For example, in the study by Gomez *et al.* [2018], 9 children aged 3 to 8 and their teachers participated in the evaluation of the game "Leo con Lula"⁸. The teachers used the game in their classes, once a day for a week. At the end of the week, the teachers answered a questionnaire to collect their feelings and experiences about the game. In the study by Golestan *et al.* [2019], evaluations were conducted with the games CARAUTI - a speech-therapy game - and BEESAUTI: a hand-eye coordination game. The evaluations aimed to verify if the children were attracted to the games, if the parents could use the games and easily interact with their children, and if the games seemed useful as therapy for ASD. To do so, 8 children with ASD and TD (aged 3 to 7 years) and their parents were recruited to play the two games at home. After using the games, the parents answered a questionnaire about each game. The questionnaires contained questions about the child's interest, the parent's difficulty in playing, ranking game graphics, among others.

In two other studies, the questionnaires were filled out by end users and stakeholders [Aruanno *et al.*, 2018; Garcia-Garcia *et al.*, 2019]. In the study by Aruanno *et al.* [2018], the evaluation involved 20 people, aged 16 to 43, with NDD, and their 3 caregivers. The evaluation was divided into sessions, in which each participant used the game for approximately 10 minutes, with the help of their caregiver. During the evaluation, an evaluator observed and made notes. At the end of each session, the participants and caregivers answered a questionnaire. The caregiver's questionnaire contained questions about device acceptability, usability of the interaction mode, task complexity, virtual assistant role. The participants' questionnaire had two questions about likability and satisfaction. In the study by Garcia-Garcia *et al.* [2019], EmoTEA, a serious game in the form of a mobile application, was evaluated. In this study, 3 children, aged 8 to 10 years old, were observed performing the following tasks: i) Browse the application; and ii) Playing the first level of the three games. The children were accompanied by an educator throughout the evaluation to help or calm them down if they tried to get up from the chair during the assessment. At the end of the evaluation, the three participants filled out the System Usability Scale (SUS) questionnaire. The educator also filled out the SUS questionnaire to provide feedback from their point of view. The questionnaire consisted of 10 questions that evaluated various aspects related to the game.

Regarding the evaluated quality properties, studies that used the *questionnaire* method alone evaluated the following set of quality properties: usability (1 study), and feasibility (3 studies). In studies where the *questionnaire* method was combined with other evaluation approaches, 6 out of the 7 identified quality properties were evaluated: usability (6), feasibility (3), impact (3), user experience (2), engagement (1), and metric efficacy (1). The *questionnaire* method was only not used in communicability evaluations (only 1 study assessed communicability).

6.5 Interview

The *interview* method was used in 11 studies, which evaluated games in the following categories: *social and socio-emotional skills* (7 studies); *cognitive skills* (2 studies); *motor skills* (1 study); and *communication skills* (1 study).

Two studies used the interview method alone, and 9 combined it with other methods. In 4 studies, it was combined with only one other method: *observation* (3 studies) and *in-game evaluation* (1 study). Another 5 studies combined the application of the *interview* method with two methods, namely: *observation* and *questionnaire* (2 studies); *in-game evaluation* and *observation* (1 study); *in-game evaluation* and *pre-post test* (1 study); *observation* and *pre-post test* (1 study).

Seven of the 11 studies that applied the *interview* method involved end-users as evaluation participants. In these studies, sample sizes ranged from 3 to 11 participants, with a mean of 7.1 participants (standard deviation of 2.7). The age range of participants ranged from 5 to 31 years. In most of these studies, participants were individuals with ASD (5 studies), followed by NDD (1 study) and ASD+TD (1 study). Four studies [Gotsis *et al.*, 2010; Giusti *et al.*, 2011; Boyd *et al.*, 2018; Jain *et al.*, 2012] interviewed both end-users (i.e., children with ASD) and stakeholders. The stakeholders involved in these studies were parents [Gotsis *et al.*, 2010; Boyd *et al.*, 2018; Jain *et al.*, 2012] and therapists [Giusti *et al.*, 2011].

For example, in the study by Jain *et al.* [2012], 9 children with ASD, aged between 5 and 12 years, were recruited to evaluate a serious game to teach facial expressions. In the session, children could play the game as long as they wanted. In the end, children and parents were interviewed about their experience with the game. Another 4 studies conducted interviews only with stakeholders [Ringland *et al.*, 2019; Carlier *et al.*, 2019; Boyd *et al.*, 2017; Mei and Guo, 2018]. Parents [Ringland *et al.*, 2019; Carlier *et al.*, 2019], teachers [Boyd *et al.*, 2017], therapists and game development students [Mei and Guo, 2018] took part in these studies.

Regarding the evaluated quality properties, studies that used the *interview* method alone evaluated user experience (1 study) and feasibility (1 study). In turn, studies where the *interview* method was combined with different approaches evaluated impact (5), usability (2), feasibility (2), and engagement (2).

6.6 Experiment

The *experiment* method was used in 9 studies, of which 3 studies evaluated games in the *social and socio-emotional skills* category; 2 studies evaluated games in the *general* category; 2 studies evaluated games in the *evaluation and measurement* category; 1 study evaluated a game in the *cognitive skills* category; and 1 study evaluated a game in the *communication skills* category.

Most of these studies (8/9) used the *experiment* as the sole evaluation method. In these 8 studies, the following quality properties were evaluated: impact (5 studies), metrics effectiveness (2 studies), and usability (1 study). Only one study combined the *experiment* with another method. In this study,

⁸Original name in Spanish.

the *experiment* method was combined with the questionnaire method to evaluate the quality property of *metrics effectiveness*.

In the 9 studies where the *experiment* method was used, the number of participants ranged from 6 to 98, with an average sample size of 25.8 participants (SD = 19.7). The age range of these participants varied from 2 to 17 years old. In most of these studies, the participants were individuals with ASD (6), followed by ASD+TD (2), and NDD (1). No stakeholders were involved in the evaluations using the *experiment* method.

6.7 Expert Review

The *expert review* method was used to evaluate games in the following categories: social and socio-emotional skills (2 studies); general (3 studies); academic skills (1 study); and evaluation and measurement (1 study). Among the 7 studies that applied the method, 4 studies used it in isolation to evaluate the following quality properties: usability (3 studies), feasibility (1 study), and communicability (1 study). One of these articles evaluated two quality properties.

Moreover, 3 studies used the *expert review* method in conjunction with other evaluation approaches. The following combinations were made with the expert review method: 1 study combined the method with a *questionnaire*; 1 study combined it with *observation*; and 1 study combined it with both - *observation* and a *questionnaire*. These studies evaluated usability (3 studies) and feasibility (2 studies). It is noteworthy that two studies evaluated two quality properties.

The studies that used the *expert review* method had the participation of experts with different profiles, namely: psychologist Thordarson and Vilhjálmsdóttir [2019]; Moura *et al.* [2016]; Sturm *et al.* [2016], lecturer in game analysis Marwecki *et al.* [2013], professional designer of games Marwecki *et al.* [2013], therapist Marwecki *et al.* [2013]; Moura *et al.* [2016], educator Marwecki *et al.* [2013]; Weilun *et al.* [2011], parents Carvalho and da Cunha [2019], and HCI expert Sousa *et al.* [2012]. Two studies that used the *expert review* method also involved end-users. In the study by Sousa *et al.* [2012], three HCI experts evaluated the usability and communicability of the WorldTour game. For this, the experts used the Heuristic Evaluation method, the Semiotic Inspection Method (SIM), and the Communicability Evaluation Method (CEM). Since the CEM is applied to evaluate the designer's communication with the user, through the interface, in real-time interaction, the study counted on the participation of two potential users - a 9-year-old girl diagnosed with pervasive developmental disorder-not otherwise specified (PDD-NOS) and a 7-year-old boy diagnosed with autism. In the study by Dragomir *et al.* [2018], three psychologists analyzed the emot-iCan game and reported their perspectives on the design of the configuration interface aimed at the game's administrator; and of the game's reward system. In addition, the psychologists provided feedback on how the players reacted to the game. Both players with typical development and with ASD participated in the test. The study did not inform the number and age range of the participants. It was only informed that the groups of participants with TD and ASD had a similar age range.

6.8 Focus Group

The *focus group* method was used in only two studies, which evaluated games in the following categories: social and socio-emotional skill and general. In the study by Loiacono *et al.* [2018], a *focus group* was conducted with therapists, and the observation method was also used to evaluate the game's usability. In the study by Soysa and Al Mahmud [2020], 20 children with ASD, aged 3 to 6 years, and 5 teachers participated in the evaluation. The children used the game, and at the end of the sessions, a focus group was conducted with the teachers to discuss usability issues with the game.

7 Threats to validity

In any systematic literature review, there are some threats regarding the validity of the results. Therefore, we seek to raise potential threats and apply strategies to try and mitigate their impacts.

First, we are aware that the the review's subject can be addressed in other areas. However, as our focus is to analyze how serious games have been evaluated from the perspective of the HCI area to assist in the future design and evaluation of games, we focused on analyzing publications in the area of HCI that are primarily in computing repositories. For this purpose, we have selected some of the main repositories and conference proceedings that store relevant works carried out in the field of Computer Science and related to serious games. Still, it is possible that existing relevant studies have not been considered. Furthermore, as the search for publications took place until March 2020, after this date, other studies adherent to the this research objective may have been published but were not included in this SLR. These new studies may present new findings, such as new evaluation methods and quality properties that are not included here.

Moreover, although we have followed the methodology and carried out a systematic process of article selection and analysis, the extraction process involves subjective interpretation and is based on the researchers' decisions. To minimize this bias, two researchers carried out each step of the process individually, as well as the data extraction, which was consolidated between them. In case of divergences or doubts, a third researcher was involved in the final decision-making.

The studies did not present all steps of their research with the same level of detail. Thus, in some steps of the analysis, inferences were made based on what was explicitly presented. For example, in the study by Sharma *et al.* [2018], the authors described that the evaluation of the Kirana game was developed in four phases. The pre- and post-evaluation took place in phases I and III, respectively. During phase IV, they observed the participants' behavior in a real situation. In phase II, the authors said the participants had played the game. Although the authors did not give further details about the participants' gaming session (phase II), they mentioned that the observer's notes were examined when presenting the analysis of their evaluation. Thus, in our analysis of the study, we assumed that the participants' sessions were observed as part of the evaluation.

In our analysis, we classified the method used according to when or how the data was collected. Although this analysis provides an overview of the methods used for evaluation, it does not take into account other aspects related to data collection for evaluation. For example, the evaluation methods grouped as *in-game evaluation* may include different data collected during the game, such as quantitative metrics (e.g., performance data) and qualitative data (e.g., the user's mood is collected through questions presented during with the game). To indicate the diversity of aspects involved in each evaluation method, we pointed them out in the discussion and presented some of the different approaches among their application. For a more detailed analysis of the evaluation methods, different aspects related to evaluation (e.g., type of data collected, moment in the design process, instrument/technology used, etc.) could be considered in the analysis. However, a challenge to do this is precisely the difference in the level of presentation of this information in the studies, because these evaluations are often presented in a short section within the article, which focuses on the development of the serious game. So, sometimes, very little (or even none) information about the evaluation methods used is presented.

8 Conclusions

This work investigated evaluation methods are being applied to assess serious games for children with ASD. Through a systematic literature review, it was possible to analyze the state-of-the-art of the literature regarding: 1) the methods used in the evaluations; 2) the set of methods used to evaluate each game category; 3) how the evaluation methods have been applied; 4) the number and profile of the participants involved in evaluations and 5) the quality properties have been considered in the games' evaluation.

Our findings indicate that there is no consolidated methodology specifically for evaluating serious games for children with ASD. Different existing methods have been used and combined in different ways to evaluate this type of game. Moreover, we did not identify a clear relationship between the game category and the methods used in its evaluation. However, it is important to note that the distribution of games in categories is very disparate (i.e., two categories account for about 58% of games), making it difficult to conduct a more significant analysis. Observation was the most used method and the one most combined with other methods. The reason for this could be that it allows the evaluator to collect data about the child's experience, without much interference in these sessions. On the other hand, the questionnaire was the only method used in evaluations for all game categories, perhaps because this method is low-cost and easy to combine with other methods.

It is known that TEA has different levels of severity [American Psychiatric Association, 2013]. Still, most of the analyzed articles do not specify the level of severity of the games' target audience. Articles usually only describe that the serious game is aimed at children with ASD or encompasses children with ASD and define the skill the game aims to improve. Future works can investigate how to take into

consideration the different levels of severity of autism in the design and use of serious games.

As presented, a large proportion of games (65/75) included representative participants of their users in the evaluation. In some situations, neurotypical individuals were also included. In these cases, the main objective was to identify differences between individuals with ASD and neurotypicals or to evaluate games aimed at stimulating interaction between children with ASD and neurotypical individuals. In addition, 19 evaluations included therapists, psychologists, teachers, or parents. These profiles represent important *stakeholders*, as some games are developed with the intention of being used as part of the children's treatment, and these *stakeholders* would be responsible for using the games with the children. In these cases, the methods that were used more frequently were questionnaire, expert review, interview, and focus group. In three evaluations, only *stakeholders* (and not users) were included.

In the evaluations, the maximum number of quality properties investigated was two. However, as they are not mutually exclusive but complementary, it would make sense to take into account the evaluation of several (or even all) of the quality properties for each game. Although, including the evaluation of several properties could be too costly, perhaps focusing on distinct properties in different times in the development or implementation of the game might be feasible and interesting. For example, evaluating the game's usability during the design of the game, and once it is ready, assess its impact. One could argue that the ultimate goal is for serious games to impact children's treatment positively. However, our findings show that the results of works that evaluated the impact of games are still incipient. They considered the impact of a particular game on a specific skill (e.g., [Sharma *et al.*, 2018; Dragomir *et al.*, 2018; Wade *et al.*, 2017]), so the results are very dependent on the context considered. Nonetheless, they suggest that games have the potential to improve the skills addressed that they focused on. Only one study reported an inconclusive evaluation and described that more tests would be needed to prove or disprove the investigated hypothesis (Hypothesis 1: The child indicates reduced levels of stress and/or anxiety after a gaming session) [Carlier *et al.*, 2019]. It is worth highlighting no study indicated any negative impacts of games on participants. At any rate, it is still necessary to advance in the investigation of the effectiveness of games, but this work can help as an initial step in this direction.

Our findings in this study indicate that the skill addressed in a game or its type are not enough to determine which method(s) would be best for their evaluation. The decision of the method or combination of methods, user profiles, number, and quality properties must take into account the objective, context, and available resources. Although this conclusion is not new to the field of HCI, it indicates the importance of having ways to support serious game project teams in their consideration of which evaluation methods would be appropriate or interesting for their context. Thus, this work brings relevant contributions to the evaluation of serious games for children with ASD, as it characterizes and discusses relevant aspects to be decided about these evaluations, based on the practice of what has been done in the field. Therefore it contributes to researchers or professionals who develop serious

games for this context, who can draw from the discussion presented in this article and use it both to expand their knowledge about how evaluation methods have been used, and as a starting point to guide their decisions regarding the evaluations that would be most interesting in their own contexts.

Based on the results of this work, future work can delve into the specific analysis either focusing on one method or on games focusing on a specific ability or generating a more detailed characterization for the focus in question, or generating support materials for these contexts (e.g. questionnaires or impact metrics used). Another interesting direction to investigate is how ethical issues are being addressed in research related to serious games for children with ASD. Additionally, future work can extend the search to other repositories (e.g. from health field), including those with an interdisciplinary focus. Finally, another relevant future work would be to extend the analysis period beyond 2020. Updating the SLR to a period after 2020 could generate complementary and new results, such as the emergence of new approaches to evaluate serious games, especially considering that the COVID-19 pandemic may have shifted or changed the focus or means of evaluation.

Declarations

Funding

This project was partially funded by the National Council of Scientific and Technological Development (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), the Innovation Center on Artificial Intelligence for Health (CIA-Saúde); and Analytical Capabilities Program - MPMG

Authors' Contributions

Ana Paula de Carvalho: Conceptualization, Methodology, Investigation, Data curation, Visualization, Validation, Writing – original draft. **Camila Santana Braz:** Investigation, Data curation, Validation, Visualization, Writing – original draft. **Raquel Oliveira Prates:** Conceptualization, Methodology, Validation, Writing – review & editing.

Availability of data and materials

The systematic literature review extracted data is available at <https://docs.google.com/spreadsheets/d/14u3skqiRQvdUA0oABwxFciHlkeq8ACoSjLJdPqSQTu0/edit?usp=sharing>

References

- Al-Hammadi, M. and Abdelazim, A. (2015). Randomness impact in digital game-based learning. In *2015 IEEE Global Engineering Education Conference (EDUCON)*, pages 806–811. DOI: 10.1109/EDUCON.2015.7096064.
- American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. American Psychiatric Association, Arlington, VA. DOI: 10.1176/appi.books.9780890425596.
- Aruanno, B., Garzotto, F., Torelli, E., and Vona, F. (2018). Hololearn: Wearable mixed reality for people with neurodevelopmental disorders (nnd). In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '18*, page 40–51, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3234695.3236351.
- Barajas, A. O., Al Osman, H., and Shirmohammadi, S. (2017). A serious game for children with autism spectrum disorder as a tool for play therapy. In *2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–7. DOI: 10.1109/SeGAH.2017.7939266.
- Bartoli, L., Garzotto, F., Gelsomini, M., Oliveto, L., and Valoriani, M. (2014). Designing and evaluating touchless playful interaction for asd children. In *Proceedings of the 2014 Conference on Interaction Design and Children, IDC '14*, page 17–26, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2593968.2593976.
- Boyd, L. E., Gupta, S., Vikmani, S. B., Gutierrez, C. M., Yang, J., Linstead, E., and Hayes, G. R. (2018). Vrsocial: Toward immersive therapeutic vr systems for children with autism. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, page 1–12, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3173574.3173778.
- Boyd, L. E., Ringland, K. E., Faucett, H., Hiniker, A., Klein, K., Patel, K., and Hayes, G. R. (2017). Evaluating an ipad game to address overselectivity in preliterate aac users with minimal verbal behavior. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '17*, page 240–249, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3132525.3132551.
- Boyd, L. E., Ringland, K. E., Haimson, O. L., Fernandez, H., Bistarkey, M., and Hayes, G. R. (2015). Evaluating a collaborative ipad game's impact on social relationships for children with autism spectrum disorder. *ACM Transactions on Accessible Computing (TACCESS)*, 7(1):1–18. DOI: 10.1145/2751564.
- Buzzi, M. C., Paolini, G., Senette, C., Buzzi, M., and Paratore, M. T. (2019). Designing an accessible web app to teach piano to students with autism. In *Proceedings of the 13th Biannual Conference of the Italian SIGCHI Chapter: Designing the next Interaction, CHIItaly '19*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3351995.3352037.
- Calderón, A. and Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers Education*, 87:396–422. DOI: <https://doi.org/10.1016/j.compedu.2015.07.011>.
- Carlier, S., Van der Paelt, S., Ongenae, F., De Backere, F., and De Turck, F. (2019). Using a serious game to reduce stress and anxiety in children with autism spectrum disorder. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth'19*, page 452–461, New

- York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3329189.3329237.
- Carreño-León, M. A., Sandoval-Bringas, J. A., Encinas, I. D., Castro, R. C., Cota, I. E., and Carrillo, A. L. (2021). Managing emotions in autistic children through serious game with tangible interfaces. In *2021 4th International Conference on Inclusive Technology and Education (CONTIE)*, pages 126–133. DOI: 10.1109/CONTIE54684.2021.00029.
- Carvalho, L. T. and da Cunha, M. X. C. (2019). Abc autism animals: An application to aid children with autism in learning (in portuguese. original title: Abc autismo animais: Um aplicativo para auxiliar a aprendizagem de crianças com autismo). *Proceedings of Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*.
- Chan, S., Cai, Y., Lu, A., Tun, N. Z., Huang, L., and Chandrasekaran, I. (2016). Virtual reality enhanced pink dolphin game for children with asd. In *Proceedings of the 3rd Asia-Europe Symposium on Simulation & Serious Gaming, VRCAI '16*, page 215–218, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3014033.3014039.
- Chen, C., Chander, A., and Uchino, K. (2019). Guided play: Digital sensing and coaching for stereotypical play behavior in children with autism. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, page 208–217, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3301275.3302309.
- Cordeiro, R. F., Ferreira, W. S., Aguiar, Y. P. C., Saraiva, J. A. G., Tardif, C., and Galy, E. (2018). The brazilian challenge to accessibility and digital inclusion for people with autistic spectrum disorders. In *Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems, IHC 2018*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3274192.3274229.
- Crovati, P., Gianotti, M., Riccardi, F., and Garzotto, F. (2019). Designing a smart toy: Guidelines from the experience with smart dolphin "sam". In *Proceedings of the 13th Biannual Conference of the Italian SIGCHI Chapter: Designing the next Interaction, CHIItaly '19*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3351995.3352041.
- Cunha, R. M. (2011). Development and evaluation of a computer game to teach vocabulary to children with autism (in portuguese. original title: Desenvolvimento e avaliação de um jogo de computador para ensino de vocabulário para crianças com autismo). *Proceedings of Games for Change*.
- Dantas, A. C., de Melo, S., Neves, L., Milessi, T., and do Nascimento, M. Z. (2020). Michelzinho: Serious game to teach emotional skills to people with autism or intellectual disorder (in portuguese. original title: Michelzinho: Jogo sério para o ensino de habilidades emocionais em pessoas com autismo ou deficiência intelectual). In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 644. DOI: 10.5753/cbie.sbie.2019.644.
- Dapogny, A., Grossard, C., Hun, S., Serret, S., Bourgeois, J., Jean-Marie, H., Foulon, P., Ding, H., Chen, L., Dubuisson, S., Grynszpan, O., Cohen, D., and Bailly, K. (2018). Jemime: A serious game to teach children with asd how to adequately produce facial expressions. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 723–730. DOI: 10.1109/FG.2018.00114.
- de Carvalho, A. P., Braz, C. S., dos Santos, S. M., Ferreira, R. A. C., and Prates, R. O. (2023). Serious games for children with autism spectrum disorder: A systematic literature review. *International Journal of Human-Computer Interaction*, 0(0):1–28. DOI: 10.1080/10447318.2023.2194051.
- de Carvalho, A. P., Braz, C. S., and Prates, R. O. (2022). How are games for autistic children being evaluated? In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems, IHC '22*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3554364.3559127.
- de Souza, C. S. (2005). *The Semiotic Engineering of Human-Computer Interaction*. The MIT Press. DOI: 10.7551/mitpress/6175.001.0001.
- Dragomir, M., Manches, A., Fletcher-Watson, S., and Pain, H. (2018). Facilitating pretend play in autistic children: Results from an augmented reality app evaluation. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '18*, page 407–409, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3234695.3241020.
- Duval, J., Rubin, Z., Segura, E. M., Friedman, N., Zlatanov, M., Yang, L., and Kurniawan, S. (2018). Spokeit: Building a mobile speech therapy experience. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '18*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3229434.3229484.
- Finkelstein, S., Barnes, T., Wartell, Z., and Suma, E. A. (2013). Evaluation of the exertion and motivation factors of a virtual reality exercise game for children with autism. In *2013 1st Workshop on Virtual and Augmented Assistive Technology (VAAT)*, pages 11–16. DOI: 10.1109/VAAT.2013.6786186.
- Finkelstein, S., Nickel, A., Barnes, T., and Suma, E. A. (2010). Astrojumper: Motivating children with autism to exercise using a vr game. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems, CHI EA '10*, page 4189–4194, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/1753846.1754124.
- Frutos, M., Bustos, I., Zapirain, B. G., and Zorrilla, A. M. (2011). Computer game to learn and enhance speech problems for children with autism. In *2011 16th International Conference on Computer Games (CGAMES)*, pages 209–216. DOI: 10.1109/CGAMES.2011.6000340.
- Garcia-Garcia, J. M., Cabañero, M. d. M., Penichet, V. M. R., and Lozano, M. D. (2019). Emotea: Teaching children with autism spectrum disorder to identify and express emotions. In *Proceedings of the XX International Conference on Human Computer Interaction, Interacción '19*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3335595.3335639.
- Garzotto, F., Gelsomini, M., Oliveto, L., and Valori-

- ani, M. (2014). Motion-based touchless interaction for asd children: A case study. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, AVI '14*, page 117–120, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2598153.2598197.
- Giacolini, L., Marti, P., and Iacono, I. (2015). Game of stimuli: An exploratory tangible interface designed for autism. In *Proceedings of the European Conference on Cognitive Ergonomics 2015, ECCE '15*, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2788412.2788444.
- Giusti, L., Zancanaro, M., Gal, E., and Weiss, P. L. T. (2011). Dimensions of collaboration on a tabletop interface for children with autism spectrum disorder. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, page 3295–3304, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/1978942.1979431.
- Glaser, N. and Schmidt, M. (2021). Systematic literature review of virtual reality intervention design patterns for individuals with autism spectrum disorders. *International Journal of Human-Computer Interaction*, 38(8):1–36. DOI: 10.1080/10447318.2021.1970433.
- Gobbo, M. R. d. M., de Barbosa, C. R. S. C., Morandini, M., and Mafort, F. (2019). Application for vocabulary gain and literacy aid aimed at children with autism spectrum disorder (in portuguese. original title: Aplicativo para ganho de vocabulário e auxílio na alfabetização destinado às crianças com transtorno do espectro autista). In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1111. DOI: 10.5753/cbie.sbie.2019.1111.
- Golestan, S., Mahmoudi-Nejad, A., and Moradi, H. (2019). A framework for easier designs: Augmented intelligence in serious games for cognitive development. *IEEE Consumer Electronics Magazine*, 8(1):19–24. DOI: 10.1109/MCE.2018.2867970.
- Gomez, J., Jaccheri, L., Torrado, J. C., and Montoro, G. (2018). Leo con lula, introducing global reading methods to children with asd. In *Proceedings of the 17th ACM Conference on Interaction Design and Children, IDC '18*, page 420–426, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3202185.3202765.
- Gotsis, M., Piggot, J., Hughes, D., and Stone, W. (2010). Smart-games: A video game intervention for children with autism spectrum disorders. In *Proceedings of the 9th International Conference on Interaction Design and Children, IDC '10*, page 194–197, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/1810543.1810569.
- Guerra, E. and Furtado, F. (2013). A proposal software for multidisciplinary treatment of autistic children. In *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6.
- Harrold, N., Tan, C. T., Rosser, D., and Leong, T. W. (2014). Copyme: A portable real-time feedback expression recognition game for children. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems, CHI EA '14*, page 1195–1200, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2559206.2581279.
- Hassan, A., Pinkwart, N., and Shafi, M. (2021). Serious games to improve social and emotional intelligence in children with autism. *Entertainment Computing*, 38:100417. DOI: 10.1016/j.entcom.2021.100417.
- Hassan, A. Z., Zahed, B. T., Zohora, F. T., Moosa, J. M., Salam, T., Rahman, M. M., Ferdous, H. S., and Ahmed, S. I. (2011). Developing the concept of money by interactive computer games for autistic children. In *2011 IEEE International Symposium on Multimedia*, pages 559–564. DOI: 10.1109/ISM.2011.99.
- Hughes, D. E., Vasquez, E., and Nicsinger, E. (2016). Improving perspective taking and empathy in children with autism spectrum disorder. In *2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–5. DOI: 10.1109/SeGAH.2016.7586232.
- Iyer, S., Mishra, R. S., Kulkarni, S. P., and Kalbande, D. (2017). Assess autism level while playing games. In *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, pages 42–47. DOI: 10.1109/CSCITA.2017.8066573.
- Jain, S., Tamersoy, B., Zhang, Y., Aggarwal, J. K., and Orvalho, V. (2012). An interactive game for teaching facial expressions to children with autism spectrum disorders. In *2012 5th International Symposium on Communications, Control and Signal Processing*, pages 1–4. DOI: 10.1109/ISCCSP.2012.6217849.
- Jouaiti, M. and Henaff, P. (2019). Robot-based motor rehabilitation in autism: a systematic review. *International Journal of Social Robotics*, 11(5):753–764. DOI: 10.1007/s12369-019-00598-9.
- Kashani-Vahid, L., Mohajeri, M., Moradi, H., and Irani, A. (2018). Effectiveness of computer games of emotion regulation on social skills of children with intellectual disability. In *2018 2nd National and 1st International Digital Games Research Conference: Trends, Technologies, and Applications (DGRC)*, pages 46–50. DOI: 10.1109/DGRC.2018.8712024.
- Khowaja, K., Banire, B., Al-Thani, D., Sqalli, M. T., Aqle, A., Shah, A., and Salim, S. S. (2020). Augmented reality for learning of children and adolescents with autism spectrum disorder (asd): A systematic review. *IEEE Access*, 8:78779–78807. DOI: 10.1109/ACCESS.2020.2986608.
- Kirst, S., Diehm, R., Bögl, K., Wilde-Etzold, S., Bach, C., Noterdaeme, M., Poustka, L., Ziegler, M., and Dziobek, I. (2022). Fostering socio-emotional competencies in children on the autism spectrum using a parent-assisted serious game: A multicenter randomized controlled trial. *Behaviour Research and Therapy*, 152:104068. DOI: <https://doi.org/10.1016/j.brat.2022.104068>.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1, Keele University, Keele, U.K.
- Koirala, A., Yu, Z., Schiltz, H., Van Hecke, A., Koth, K. A., and Zheng, Z. (2019). An exploration of using virtual reality to assess the sensory abnormalities in children

- with autism spectrum disorder. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 293–300, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3311927.3323118.
- Kořakowska, A., Landowska, A., and Karpjenko, K. (2017). Gyroscope-based game revealing progress of children with autism. In *Proceedings of the 2017 International Conference on Machine Learning and Soft Computing*, ICMLSC '17, page 19–24, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3036290.3036324.
- Kousar, S., Mehmood, N., and Ahmed, S. (2019). Serious games for autism children: A comparative study. *University of Sindh Journal of Information and Communication Technology*, 3(3):162–170.
- Kurniawati, A., Kusumaningsih, A., and Hasan, I. (2019). Class vr: Learning class environment for special educational needs using virtual reality games. In *2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pages 1–5. DOI: 10.1109/CENIM48368.2019.8973353.
- Li, B., Atyabi, A., Kim, M., Barney, E., Ahn, A. Y., Luo, Y., Aubertine, M., Corrigan, S., St. John, T., Wang, Q., Mademtzi, M., Best, M., and Shic, F. (2018). Social influences on executive functioning in autism: Design of a mobile gaming platform. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3173574.3174017.
- Loiacono, T., Trabucchi, M., Messina, N., Matarazzo, V., Garzotto, F., and Beccalupa, E. A. (2018). Social matchup -: A memory-like virtual reality game for the enhancement of social skills in children with neurodevelopmental disorders. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3170427.3188525.
- Maenner, M. J., Shaw, K. A., Bakian, A. V., Bilder, D. A., Durkin, M. S., Esler, A., Furnier, S. M., Hallas, L., Hall-Lande, J., Hudson, A., Hughes, M. M., Patrick, M., Pierce, K., Poynter, J. N., Salinas, A., Shenouda, J., Vehorn, A., Warren, Z., Constantino, J. N., DiRienzo, M., Fitzgerald, R. T., Grzybowski, A., Spivey, M. H., Pettygrove, S., Zahorodny, W., Ali, A., Andrews, J. G., Baroud, T., Gutierrez, J., Hewitt, A., Lee, L.-C., Lopez, M., Mancilla, K. C., McArthur, D., Schwenk, Y. D., Washington, A., Williams, S., and Cogswell, M. E. (2021). Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2018. *Morbidity and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002)*, 70(11):1–16. DOI: 10.15585/mmwr.ss7011a1.
- Marchi, E., Schuller, B., Baird, A., Baron-Cohen, S., Lasalle, A., O'Reilly, H., Pigat, D., Robinson, P., Davies, I., Baltrušaitis, T., Adams, A., Mahmoud, M., Golan, O., Fridenson-Hayo, S., Tal, S., Newman, S., Meir-Goren, N., Camurri, A., Piana, S., Bölte, S., Sezgin, M., Alyuz, N., Rynkiewicz, A., and Baranger, A. (2019). The asc-inclusion perceptual serious gaming platform for autistic children. *IEEE Transactions on Games*, 11(4):328–339. DOI: 10.1109/TG.2018.2864640.
- Marques, A. B. and Monte, L. d. S. (2022). How are software technologies being evaluated with autistic users? a systematic mapping. *Universal Access in the Information Society*, 21:587–597. DOI: 10.1007/s10209-021-00794-3.
- Marwecki, S., Rädle, R., and Reiterer, H. (2013). Encouraging collaboration in hybrid therapy games for autistic children. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 469–474, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2468356.2468439.
- Mei, C. and Guo, R. (2018). Enable an innovative prolonged exposure therapy of attention deficits on autism spectrum through adaptive virtual environments. In *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–4. DOI: 10.1109/VS-Games.2018.8493421.
- Mei, C., Zahed, B. T., Mason, L., and Ouarles, J. (2018). Towards joint attention training for children with asd - a vr game approach and eye gaze exploration. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 289–296. DOI: 10.1109/VR.2018.8446242.
- Mir, H. Y. and Khosla, A. K. (2018). Kinect based game for improvement of sensory, motor and learning skills in autistic children. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1670–1674. DOI: 10.1109/ICCONS.2018.8662894.
- Moura, D., de Oliveira Filh, D. L. S., Laertius, D., Silva, A. J. G., Paiva, P., de Sales, T., Cavalcante, R., and Queiroz, F. (2016). Teo: An interactive game suite to support the treatment of children with autism (in portuguese. original title: Teo: Uma suíte de jogos interativos para apoio ao tratamento de crianças com autismo). In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 27, page 627. DOI: 10.5753/cbie.sbie.2016.627.
- Neto, O. P. d. S., de Sousa, V. H. V., Batista, G. B., Santana, F. C. B. G., and Junior, J. M. B. O. (2013). G-tea: A tool to support learning for children with autism spectrum disorder, based on the aba methodology (in portuguese. original title: G-tea: Uma ferramenta no auxílio da aprendizagem de crianças com transtorno do espectro autista, baseada na metodologia aba). In *Proceedings of Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pages 16–18.
- Noor, H. A. M., Shahbodan, F., and Pee, N. C. (2012). Serious game for autism children: review of literature. *World Academy of Science, Engineering and Technology*, 6(4):554–559. DOI: 10.5281/zenodo.1333272.
- Parisa Ghanouni, Tal Jarus, J. G. Z. and Lucyshyn, J. (2021). An interactive serious game to target perspective taking skills among children with asd: A usability testing. *Behaviour & Information Technology*, 40(16):1716–1726. DOI: 10.1080/0144929X.2020.1776770.
- Petri, G. and Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? a system-

- atic literature review. *Computers Education*, 107:68–90. DOI: <https://doi.org/10.1016/j.compedu.2017.01.004>.
- Piana, S., Malagoli, C., Usai, M. C., and Camurri, A. (2019). Effects of computerized emotional training on children with high functioning autism. *IEEE Transactions on Affective Computing*, 12(4):1045–1054. DOI: 10.1007/s10803-019-04135-5.
- Pistoljevic, N. and Hulusic, V. (2017). An interactive e-book with an educational game for children with developmental disorders: A pilot user study. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 87–93. DOI: 10.1109/VS-GAMES.2017.8056575.
- Porayska-Pomsta, K., Alcorn, A. M., Avramides, K., Beale, S., Bernardini, S., Foster, M. E., Frauenberger, C., Good, J., Guldberg, K., Keay-Bright, W., Kossyvakis, L., Lemon, O., Mademtzi, M., Menzies, R., Pain, H., Rajendran, G., Waller, A., Wass, S., and Smith, T. J. (2018). Blending human and artificial intelligence to support autistic children's social communication skills. *ACM Trans. Comput.-Hum. Interact.*, 25(6):1–35. DOI: 10.1145/3271484.
- Rahman, M. M., Ferdous, S., and Ahmed, S. I. (2010). Increasing intelligibility in the speech of the autistic children by an interactive computer game. In *2010 IEEE International Symposium on Multimedia*, pages 383–387. DOI: 10.1109/ISM.2010.64.
- Rapela, J., Lin, T.-Y., Westerfield, M., Jung, T.-P., and Townsend, J. (2012). Assisting autistic children with wireless eeg technology. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3504–3506. DOI: 10.1109/EMBC.2012.6346721.
- Ribeiro, P. C., Baere Pederassi Lomba de Araujo, B., and Raposo, A. (2014). Comfim: A cooperative serious game to encourage the development of communicative skills between children with autism. In *2014 Brazilian Symposium on Computer Games and Digital Entertainment*, pages 148–157. DOI: 10.1109/SBGAMES.2014.19.
- Ringland, K. E., Wolf, C. T., Boyd, L., Brown, J. K., Palermo, A., Lakes, K., and Hayes, G. R. (2019). Dancecraft: A whole-body interactive system for children with autism. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 572–574, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3308561.3354604.
- Ritterfeld, U., Cody, M., and Vorderer, P. (2009). *Serious games: Mechanisms and effects*. Routledge. DOI: 10.4324/9780203891650.
- Rodrigues, J. H., Silva, L., and Bellon, O. R. P. (2018). Transferring facial expression to animated avatars: Assisting children with autism spectrum disorder (in portuguese. original title: Transferência de expressão facial para avatares animados: Auxiliando crianças com transtorno do espectro autista). *XVII Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*.
- Rouhi, A., Spitale, M., Catania, F., Cosentino, G., Gelsomini, M., and Garzotto, F. (2019). Emotify: Emotional game for children with autism spectrum disorder based-on machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion, IUI '19*, page 31–32, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3308557.3308688.
- Sharma, S., Varkey, B., Achary, K., Hakulinen, J., Turunen, M., Heimonen, T., Srivastava, S., and Rajput, N. (2018). Designing gesture-based applications for individuals with developmental disabilities: guidelines from user studies in india. *ACM Transactions on Accessible Computing (TACCESS)*, 11(1):1–27. DOI: 10.1145/3161710.
- Silva, G. F. M. and Raposo, A. B. (2016). Identifying awareness requirements in face-to-face collaborative applications for users with autism spectrum disorders. In *Anais do XIII Simpósio Brasileiro de Sistemas Colaborativos*, pages 61–75. SBC. DOI: 10.5753/sbsc.2016.9492.
- Silva, G. M., Souto, J. J. d. S., Fernandes, T. P., Bolis, I., and Santos, N. A. (2021). Interventions with serious games and entertainment games in autism spectrum disorder: A systematic review. *Developmental Neuropsychology*, 46(7):463–485. PMID: 34595981. DOI: 10.1080/87565641.2021.1981905.
- Silva-Calpa, G. F. M., Raposo, A. B., and Suplino, M. (2018). Coasd: A tabletop game to support the collaborative work of users with autism spectrum disorder. In *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–8. DOI: 10.1109/SeGAH.2018.8401358.
- Sousa, F. R. M., Costa, E. A. B., and de Castro, T. H. C. (2012). Worldtour: Software to support teaching of autistic children (in portuguese. original title: Worldtour: Software para suporte no ensino de crianças autistas). In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 23.
- Soysa, A. I. and Al Mahmud, A. (2020). Tangible play and children with asd in low-resource countries: A case study. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '20, page 219–225, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3374920.3374951.
- Spitale, M., Gelsomini, M., Beccaluva, E., Viola, L., and Garzotto, F. (2019). Meeting the needs of people with neuro-developmental disorder through a phygital approach. In *Proceedings of the 13th Biannual Conference of the Italian SIGCHI Chapter: Designing the next Interaction*, CHIItaly '19, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3351995.3352055.
- Sturm, D., Peppe, E., and Ploog, B. (2016). emotican: Design of an assessment game for emotion recognition in players with autism. In *2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–7. DOI: 10.1109/SeGAH.2016.7586228.
- Sukiennik, R., Marchezan, J., and Scornavacca, F. (2021). Challenges on diagnosing autism spectrum disorder in brazil: a country as big as the social differences it presents. *Frontiers in Neurology*, 12:1010. DOI: 10.3389/fneur.2021.598073.

- Susi, T., Johannesson, M., and Backlund, P. (2007). Serious games : An overview. Technical Report HS-IKI-TR-07-001, University of Skövde, School of Humanities and Informatics.
- Thordarson, A. and Vilhjálmsón, H. H. (2019). Socuevr: Virtual reality game for social cue detection training. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA '19*, page 46–48, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3308532.3329440.
- Tsikinas, S. and Xinogalos, S. (2019). Studying the effects of computer serious games on people with intellectual disabilities or autism spectrum disorder: A systematic literature review. *Journal of Computer Assisted Learning*, 35(1):61–73. DOI: 10.1111/jcal.12311.
- Tsikinas, S., Xinogalos, S., and Satratzemi, M. (2016). Review on serious games for people with intellectual disabilities and autism. *Proceedings of the European Conference on Games-based Learning*, 2016-January:696–703.
- Uzuegbunam, N., Wong, W.-H., Cheung, S.-c. S., and Ruble, L. (2015). Mebook: Kinect-based self-modeling intervention for children with autism. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. DOI: 10.1109/ICME.2015.7177518.
- Vallefuoco, E., Bravaccio, C., Gison, G., and Pepino, A. (2021). Design of a serious game for enhancing money use in teens with autism spectrum disorder. In De Paolis, L. T., Arpaia, P., and Bourdot, P., editors, *Augmented Reality, Virtual Reality, and Computer Graphics*, pages 339–347, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-87595-4_25.
- Virnes, M., Kärnä, E., and Vellonen, V. (2015). Review of Research on Children with Autism Spectrum Disorder and the Use of Technology. *Journal of Special Education Technology*, 30(1):13–27. DOI: 10.1177/016264341503000102.
- Wade, J., Sarkar, A., Swanson, A., Weitlauf, A., Warren, Z., and Sarkar, N. (2017). Process measures of dyadic collaborative interaction for social skills intervention in individuals with autism spectrum disorders. *ACM Transactions on Accessible Computing (TACCESS)*, 10(4):1–19. DOI: 10.1145/3107925.
- Weilun, L., Elara, M. R., and Garcia, E. M. A. (2011). Virtual game approach for rehabilitation in autistic children. In *2011 8th International Conference on Information, Communications Signal Processing*, pages 1–6. DOI: 10.1109/ICICS.2011.6174256.
- Xianmei, L. (2017). A review of somatic games intervention for children with autism spectrum disorders. *Journal of Exceptional People*, 2(11):83. DOI: 10.5281/zenodo.1333272.
- Yanez-Gomez, R., Cascado-Caballero, D., and Sevillano, J.-L. (2017). Academic methods for usability evaluation of serious games: a systematic review. *Multimedia Tools and Applications*, 76(4):5755–5784. DOI: 10.1007/s11042-016-3845-9.
- Zakari, H. M., Ma, M., and Simmons, D. (2014). A review of serious games for children with autism spectrum disorders (asd). In Ma, M., Oliveira, M. F., and Baalsrud Hauge, J., editors, *Serious Games Development and Applications*, pages 93–106, Cham. Springer International Publishing. DOI: 10.1007/978-3-319-11623-5_9.
- Zhang, L., Fu, Q., Swanson, A., Weitlauf, A., Warren, Z., and Sarkar, N. (2018). Design and evaluation of a collaborative virtual environment (comove) for autism spectrum disorder intervention. *ACM Transactions on Accessible Computing (TACCESS)*, 11(2):1–22. DOI: 10.1145/3209687.
- Zhao, H., Swanson, A. R., Weitlauf, A. S., Warren, Z. E., and Sarkar, N. (2018). Hand-in-hand: A communication-enhancement collaborative virtual reality system for promoting social interaction in children with autism spectrum disorders. *IEEE transactions on human-machine systems*, 48(2):136–148. DOI: 10.1109/THMS.2018.2791562.