Contents

Original Papers

Digital Interventions for Stress Among Frontline Health Care Workers: Results From a Pilot Feasibility Cohort Trial (e42813)
Caroline Espinola, Binh Nguyen, Andrei Torres, Walter Sim, Alice Rueda, Lindsay Beavers, Douglas Campbell, Hyejung Jung, Wendy Lou, Bill Kapralos, Elizabeth Peter, Adam Dubrowski, Sridhar Krishnan, Venkat Bhat. 3

A Serious Game to Train Rhythmic Abilities in Children With Dyslexia: Feasibility and Usability Study (e42733)
Francois Vonthron, Antoine Yuen, Hugues Pellerin, David Cohen, Charline Grossard. 19

A Novel Casual Video Game With Simple Mental Health and Well-Being Concepts (Match Emoji): Mixed Methods Feasibility Study (e46697)
Russell Pine, James Mbinta, Lisa Te Morenga, Theresa Fleming. 35

A Smartphone-Gamified Virtual Reality Exposure Therapy Augmented With Biofeedback for Ailurophobia: Development and Evaluation Study (e34535)
Ali Khaleghi, Abbas Narimani, Zahra Aghaie, Anahita Khorrani Banaraki, Peyman Hassani-Abharian. 44

Exploring the Use of a Learning-Based Exergame to Enhance Physical Literacy, Soft Skills, and Academic Learning in School-Age Children: Pilot Interventional Study (e53072)
Aurelie Goncalves, Florence Lesplau, Gaëtan Briet, Eugénie Vaillant-Coindard, Angèle Palermo, Elsa Decobert, Nathan Allegret-Bourdon, Elodie Charbonnier. 92

Design of Virtual Reality Exergames for Upper Limb Stroke Rehabilitation Following Iterative Design Methods: Usability Study (e48900)
Julian Villada Castillo, Maria Montoya Vega, John Muñoz Cardona, David Lopez, Leonardo Quiñones, Oscar Henao Gallo, Jose Lopez. 107

A Serious Game ("Fight With Virus") for Preventing COVID-19 Health Rumors: Development and Experimental Study (e45546)
Shuo Xiong, Long Zuo, Qiwei Chen, Zhang Zeliang, Mohd Nor Akmal Khalid. 121

Reviews

The Role of AI in Serious Games and Gamification for Health: Scoping Review (e48258)
Daniel Tolks, Johannes Schmidt, Sebastian Kuhn. 161
The Effects of Serious Games on Cardiopulmonary Resuscitation Training and Education: Systematic Review With Meta-Analysis of Randomized Controlled Trials (e52990)
Pengfei Cheng, Yangxi Huang, Pengyu Yang, Haizhen Wang, Baichao Xu, Chaoran Qu, Hua Zhang. ................................. 175

Electronic Interactive Games for Glycemic Control in Individuals With Diabetes: Systematic Review and Meta-Analysis (e43574)
WenQi Yao, YiBing Han, Li Yang, Ying Chen, ShengZhe Yan, YanZhen Cheng. .............................................................. 188

Virtual Reality Therapy for the Management of Chronic Spinal Pain: Systematic Review and Meta-Analysis (e50089)
Tongtong Zhang, Xin Li, Xuan Zhou, Lixia Zhan, Fan Wu, Zetan Huang, Yuxun Sun, Yufei Feng, Qing Du. ........................ 198

Extended Reality–Based Mobile App Solutions for the Therapy of Children With Autism Spectrum Disorders: Systematic Literature Review (e49906)
Marian-Vladut Toma, Cristina Turcu, Corneliu Turcu, Sorin Vlad, Doru Tilute, Paul Pascu. .......................................................... 218

Immersive Virtual Reality–Based Methods for Assessing Executive Functioning: Systematic Review (e50282)
Rebecca Kirkham, Lars Kooijman, Lucy Albertella, Dan Myles, Murat Yücel, Kristian Rotaru. ................................................ 243

Digital Gamification Tools to Enhance Vaccine Uptake: Scoping Review (e47257)
Hina Hakim, S Driedger, Dominique Gagnon, Julien Chevrier, Geneviève Roch, Eve Dubé, Holly Witteman. .......................... 269

Research Letter

Health Care Professional–Supported Co-Design of a Mime Therapy–Based Serious Game for Facial Rehabilitation (e52661)
Daniel Sousa, Silmar Teixeira, José Fontenele, Renato Santos, Leynilson Pereira, Rodrigo Baluz, Victor Bastos, Ariel Teles. ............... 286
Digital Interventions for Stress Among Frontline Health Care Workers: Results From a Pilot Feasibility Cohort Trial

Caroline W Espinola1,2*, MSc, MD; Binh Nguyen3*, BSc; Andrei Torres4*, BArch, MEng; Walter Sim2, BSc; Alice Rueda3, PhD; Lindsay Beavers5,6, BSc, MPT; Douglas M Campbell5,7,8,9, MSc, MD; Hyejung Jung10, MSc; Wendy Lou10, PhD; Bill Kapralos1, PhD; Elizabeth Peter11, PhD; Adam Dubrowski4, PhD; Sridhar Krishnan3, PhD; Venkat Bhat1,2, MSc, MD

1Department of Psychiatry, University of Toronto, Toronto, ON, Canada
2Interventional Psychiatry Program, St. Michael’s Hospital, Unity Health Toronto, Toronto, ON, Canada
3Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada
4maxSIMhealth Group, Ontario Tech University, Oshawa, ON, Canada
5Allan Waters Family Simulation Program, Unity Health Toronto, Toronto, ON, Canada
6Department of Physical Therapy, University of Toronto, Toronto, ON, Canada
7Neonatal Intensive Care Unit, St. Michael’s Hospital, Unity Health Toronto, Toronto, ON, Canada
8Li Ka Shing Knowledge Institute, St. Michael’s Hospital, Unity Health Toronto, Toronto, ON, Canada
9Department of Pediatrics, Faculty of Medicine, University of Toronto, Toronto, ON, Canada
10Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
11Lawrence S. Bloomberg, Faculty of Nursing, University of Toronto, Toronto, ON, Canada

* these authors contributed equally

Corresponding Author:
Venkat Bhat, MSc, MD
Department of Psychiatry
University of Toronto
250 College Street, 8th floor
Toronto, ON
Canada
Phone: 1 416 360 4000 ext 76404
Email: venkat.bhat@utoronto.ca

Abstract

Background: The COVID-19 pandemic has challenged the mental health of health care workers, increasing the rates of stress, moral distress (MD), and moral injury (MI). Virtual reality (VR) is a useful tool for studying MD and MI because it can effectively elicit psychophysiological responses, is customizable, and permits the controlled study of participants in real time.

Objective: This study aims to investigate the feasibility of using an intervention comprising a VR scenario and an educational video to examine MD among health care workers during the COVID-19 pandemic and to use our mobile app for longitudinal monitoring of stress, MD, and MI after the intervention.

Methods: We recruited 15 participants for a compound intervention consisting of a VR scenario followed by an educational video and a repetition of the VR scenario. The scenario portrayed a morally challenging situation related to a shortage of life-saving equipment. Physiological signals and scores of the Moral Injury Outcome Scale (MIOS) and Perceived Stress Scale (PSS) were collected. Participants underwent a debriefing session to provide their impressions of the intervention, and content analysis was performed on the sessions. Participants were also instructed to use a mobile app for 8 weeks after the intervention to monitor stress, MD, and mental health symptoms. We conducted Wilcoxon signed rank tests on the PSS and MIOS scores to investigate whether the VR scenario could induce stress and MD. We also evaluated user experience and the sense of presence after the intervention through semi-open-ended feedback and the Igroup Presence Questionnaire, respectively. Qualitative feedback was summarized and categorized to offer an experiential perspective.

Results: All participants completed the intervention. Mean pre- and postintervention scores were respectively 10.4 (SD 9.9) and 13.5 (SD 9.1) for the MIOS and 17.3 (SD 7.5) and 19.1 (SD 8.1) for the PSS. Statistical analyses revealed no significant pre-
to postintervention difference in the MIOS and PSS scores ($P=.11$ and $P=.22$, respectively), suggesting that the experiment did not acutely induce significant levels of stress or MD. However, content analysis revealed feelings of guilt, shame, and betrayal, which relate to the experience of MD. On the basis of the Igroup Presence Questionnaire results, the VR scenario achieved an above-average degree of overall presence, spatial presence, and involvement, and slightly below-average realism. Of the 15 participants, 8 (53%) did not answer symptom surveys on the mobile app.

**Conclusions:** Our study demonstrated VR to be a feasible method to simulate morally challenging situations and elicit genuine responses associated with MD with high acceptability and tolerability. Future research could better define the efficacy of VR in examining stress, MD, and MI both acutely and in the longer term. An improved participant strategy for mobile data capture is needed for future studies.

**International Registered Report Identifier (IRRID):** RR2-10.2196/32240

**KEYWORDS**
virtual reality; simulation; mobile app; stress; moral distress; moral injury; COVID-19; mobile phone

### Introduction

#### Background

The COVID-19 pandemic has exerted unprecedented strain on health care workers (HCWs) globally [1]. Frontline HCWs have been forced to make difficult medical decisions that are contrary to their moral and professional principles and to work in conditions where they cannot meet standards of quality care [2,3], which has put them at a greater risk of experiencing moral distress (MD) than possibly ever before [4,5]. Distressing situations such as being forced to deal with a shortage of personal protective equipment and having to prioritize who will receive life-sustaining treatment have become common during the pandemic. For HCWs, experiencing such situations may cause significant emotional burden and induce the phenomenon of MD [6-8]. MD is defined as distress stemming from the inability to enact actions believed to be morally right owing to external constraints [8,9]. Moral injury (MI), an extreme form of MD, can occur when individuals witness or perpetrate actions that violate deeply held moral beliefs, resulting in severe emotional reactions with long-lasting consequences [7]. However, further investigation is needed to enable a more precise distinction between MD and MI [7].

The first description of MI was made in the military context by Shay [10] and was defined as a betrayal of moral character, usually as a result of the actions of a person in a position of authority [10], leading to feelings of powerlessness, helplessness, and loss of faith in humanity [7,10]. Shay [11] argues that MI occurs when the following conditions are met: (1) there has been a betrayal of what is considered right (2) by someone holding legitimate authority and (3) in high-stakes situations. Litz et al [12] expanded the concept of MI to include “the lasting psychological, biological, spiritual, behavioral, and social impact of perpetrated, failing to prevent, or bearing witness to acts that transgress deeply held moral beliefs and expectations.” As part of the definition, the authors also defined potentially morally injurious events (PMIEs) as the acts of perpetrating, failing to prevent harm, or bearing witness to acts that transgress deeply held moral beliefs [12]. Experiencing a PMIE is frequently associated with feelings of betrayal, guilt, shame, and self-blame [13]. Furthermore, PMIEs may not only cause acute MD but can also have long-term consequences because MD and MI may develop weeks or months after a PMIE [14].

MI was originally associated with, and frequently co-occurs with, posttraumatic stress disorder (PTSD) [13], which has been conceptualized as a fear-related disorder [15,16]. However, MI has not yet been defined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [16], and a PMIE does not necessarily fulfill the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition criterion A for PTSD. The concept of MI was conceived to encompass the following criteria, among others: (1) reexperiencing self-referential moral emotions (eg, anger, guilt, and shame); (2) strong negative beliefs about the self, the world, and others; and (3) self-destructive behaviors that inflict severe distress or functional impairment [17,18]. In addition, emerging literature has defined MI as being mechanistically different from PTSD [13,15]. A positron emission tomography study in veterans with PTSD showed that regional blood glucose metabolism differed according to the nature of traumatic exposure as follows: the group with PTSD owing to danger-based traumas (ie, life-threatening events) showed higher metabolism in the amygdala; by contrast, the group with PTSD secondary to non–danger-based traumas (eg, MI by self or others) had increased metabolism in the precuneus [19], a region that has been associated with the processing of self-referential feelings (eg, shame and guilt) [15]. Therefore, further research is needed to determine the ecological validity of MI as an independent diagnostic category [13]. In addition, there is a need to investigate specific interventions for MI because it has been found to not generally respond to evidence-based treatments for PTSD [12,17]; for example, moral resilience training, the development of emotional intelligence skills, and strategies for promoting moral repair have already been proposed as specific treatments for MI and are currently under investigation [5,17,20].

Although MI has been largely studied in military contexts [17], it is also applicable to HCWs, particularly in light of the COVID-19 pandemic. However, MI and PMIEs are poorly...
understood in this context. Čartolovni et al [7] argue that MI occurs in HCs when they experience PMIEs involving high-stakes situations that are beyond their control. To investigate MI in the COVID-19 context, Rushon et al [5] conducted a survey with frontline HCs and reported an overall prevalence rate of 32% for MI, with nurses being the most affected (38%). Fewer years of experience were positively associated with MI, whereas religious affiliation or spirituality and higher levels of moral resilience were associated with lower MI scores. In addition, the study showed a moderate correlation between MI and various ethically challenging situations, such as experiencing negative consequences at work after expressing safety concerns, working with limited resources, and carrying out decisions of others which threaten one’s own values [5].

Litam and Balkin [4] examined the relationship between MI and the professional quality of life in a convenience sample of HCs during the COVID-19 pandemic. The authors reported that secondary traumatic stress was a strong predictor of MI in frontline HCs, but the contribution of compassion, satisfaction, and burnout to MI scores was nonsignificant. Of note, nurses had significantly higher burnout scores than physicians. Zerach and Levi-Belz [21] conducted a survey to investigate the patterns of exposure to PMIEs in a sample of HCs and social care workers during the COVID-19 pandemic. The prevalence rate of symptoms of MI was 40%, with betrayal events being the most frequent PMIEs with a prevalence rate of 62%. In general, exposure to PMIEs was positively related to perceived stress, depression, anxiety, and self-criticism, whereas it was negatively associated with self-compassion. Interestingly, the duration of care for patients with COVID-19 was not associated with MI [21].

To increase the ecological validity of MI as a diagnostic entity, the experiences of the MD-MI continuum should be examined using accurate methods [13]. To date, several measurement instruments have been developed to identify MI outside of military contexts, including the Moral Injury Symptom Scale–Healthcare Professionals version [22] and the Moral Injury Outcome Scale (MIOS) [18]. The MIOS is a self-rated scale, developed as an assessment tool to evaluate MI as a multidimensional outcome [18]. This scale comprises 10 binary (yes or no) questions and 15 five-point Likert scale questions about experiencing a PMIE and feelings associated with this event; higher scores indicate greater severity of MI symptoms. At the end, the MIOS has an additional 7-point Likert scale question that assesses the extent to which the experience of PMIEs has interfered in one’s self-care or caused functional impairment (from not at all to extremely). The MIOS is in the final stage of development by the MIOS Consortium [18].

Conducting interventional studies to investigate the impact of PMIEs on mental health in real-world settings is challenging owing to operational constraints. This is especially true in health care, where limitations imposed by patient privacy regulations may make traditional clinical trials in MI impractical. Another important aspect to consider is the ethical implications of submitting an already strained workforce to moral stressors in an uncontrolled real-world environment such as an intensive care unit (ICU). A promising strategy to address these limitations is the use of virtual reality (VR) scenarios. VR is a powerful technology for examining mental health and the MD-MI continuum because it offers several advantages over traditional observational research in naturalistic environments. First, VR allows researchers to observe, monitor, and potentially support participants in fully controlled environments in real time [23]; therefore, it is safer and provides more accurate measures of one’s reactions to ethically challenging situations compared with observational studies in naturalistic environments. Second, VR allows for the design of fully customizable scenarios [23], making it especially suitable to simulate real-world scenarios in health care that otherwise would be impractical to replicate. As traumatic events in both PTSD and MI are highly idiosyncratic, and treatment for PTSD requires exposure to individual cues, we assume that virtual environment customization should be a critical feature to provide personalized and effective interventions to treat MD and MI [24]. In addition, extensive evidence has demonstrated the effectiveness of VR-based interventions for PTSD [25-27]. Third, VR environments can effectively elicit real psychophysiological responses because individuals are immersed in virtual scenarios as if these were real events, with the advantage of enabling real-time data capture [23,24]. All these advantages make VR-based trials ideal to study the MD-MI phenomena in HCs. However, no prior research has investigated the feasibility of VR interventions to examine MD and MI in the context of the COVID-19 pandemic.

Objectives

The overarching goal of this study was to determine the feasibility of using a compound VR intervention to examine MD and MI among HCs during the COVID-19 pandemic. To achieve this, we designed a VR scenario in which HCs faced a morally challenging situation in a midpandemic hospital environment while being monitored for acute psychological and physiological measures of stress. As outlined in our protocol paper [28], our aims were to (1) evaluate the feasibility of using a VR scenario to simulate the experience of a COVID-19–related morally challenging event by using measures of tolerability, dropouts, and suitability of the virtual scenario; (2) assess the potential of our VR scenario to elicit mild stress and MD, as measured by quantitative self-report questionnaires as well as qualitative analyses of semistructured interviews; and (3) investigate the feasibility of our novel mobile app (DiG App) for longitudinal monitoring of stress and MD in naturalistic settings in the 8 weeks after the intervention.

On the basis of the findings with PTSD [25-27], we hypothesized that VR scenarios would be a feasible method for assessing MD and MI. Given the ability of VR to generate genuine responses, we additionally hypothesized that our virtual scenario would significantly increase stress levels and elicit feelings and symptoms associated with MD and MI. Finally, we hypothesized that our mobile app would successfully capture symptoms associated with stress and MD in the 8-week follow-up.

To the best of our knowledge, this pilot study is the first to assess the feasibility of using a VR scenario to simulate the experience of a morally challenging event related to the
COVID-19 pandemic by HCWs while assessing its acute perceptual, psychological, and physiological effects in real time.

Methods

Study Design

In this single-cohort pilot study (ClinicalTrials.gov: NCT05001542), we adopted a multimethod approach in a pretest-posttest design to develop a compound intervention consisting of three successive parts: (1) a VR scenario to simulate a morally complex situation, (2) an educational video on MI and appropriate mitigation strategies, and (3) a repetition of the VR scenario. The intervention was followed by longitudinal data collection of mental health and MI surveys using our mobile app. The MI educational video was based on the Center of Excellence on PTSD guide [29] that summarized the causes and identifiers of MI and potential interventions to mitigate MD. The effectiveness of the VR-based educational intervention was assessed using the MIOS [18], the Perceived Stress Scale (PSS) [30], and the Igroup Presence Questionnaire (IPQ) [31]. The PSS is a self-reported measure of stress, whereas the IPQ evaluates the experience of presence during the VR scenario. As previously mentioned, the MIOS is a self-rated scale that was developed as an assessment tool to evaluate MI. For the purposes of this pilot study, we adopted a brief version of the MIOS (hereinafter referred to as the MIOS), which comprises 10 five-point Likert scale questions and 4 binary (yes or no) questions [32]. During the VR scenario, respiratory impedance, electrocardiography (ECG), galvanic skin response, and photoplethysmography were continuously collected. In addition to the original signals, we extracted the derivation of these signals, including ECG pulse rate, ECG RR interval, respiratory rate, and elevated respiratory rate. A visualization of the VR experimental flow can be seen in Figure 1. Further details on the intervention and data collection have been explained and outlined in the paper by Nguyen et al [28].

Figure 1. Flowchart of the virtual reality (VR) experiment. MIOS: Moral Injury Outcome Scale; PTSD: posttraumatic stress disorder.
The experimental session was divided into prebrief, preintervention test, intervention video, postintervention test, and debrief components (Figure 1). The preintervention test and postintervention test were conducted in VR, whereas the prebrief and debrief occurred outside the virtual environment. The MIOS was performed at 4 time points as follows: as a paper-based version for the prebrief and debrief and in the virtual scenario for the preintervention test and postintervention test. The PSS was performed twice, at prebrief and debrief. The MIOS and the PSS focus on symptoms of MD and stress, respectively, over the last month. However, when answering these scales, participants were told to rate symptoms at that exact moment. The goal of the prebrief was to explain how the physiological data would be collected and prepare the participant for the VR scenario; it consisted of an orientation to the virtual space and equipment, safety precautions, and the expected outcome of the study. During the preintervention test, participants were immersed in the VR scenario where they took on the role of a physician in an ICU during the COVID-19 pandemic. To experience the VR scenario, participants used a VR headset and 2 wireless controllers that tracked their head and hand movements, mapping it to an avatar. Semitranslucent panels were displayed as spatial elements in the VR scenario (Figure 2), providing information to the participant in the form of the dialogue panel (which displayed the current nonplayable character’s photograph, name, and the text version of the dialogue being spoken) and the interaction panel (which displayed a list of available choices and responses for the participant to choose from).

In the scenario, a shortage of life-saving equipment resulted in the decision to move a ventilator from 1 patient to another patient who had a greater chance of survival. After being informed of this, the participant’s avatar appeared in the next scene, where they had to communicate this decision to the first patient’s family and respond to the family’s reactions of frustration and anger. After completing the preintervention test and while still immersed in the VR scenario, participants watched a brief 2D educational video comprising key concepts of MD and MI and adaptive behaviors to cope with morally complex situations at the individual, team, and organizational levels. Participants then completed the postintervention test, which consisted in a repetition of the VR scenario played in the preintervention test. Finally, in the debrief, participants were asked open-ended questions to encourage them to describe their experiences in the virtual setting, followed by an exit survey [28].

After the experiment, participants were instructed to use our mobile app [33] to collect passive and active data for distress monitoring during the following 8 weeks. As MI may have a delayed onset, such data collection allows for longer-term monitoring of emotions associated with MD, offering insights into the distress experienced in real time.

**Participants**

Participants were recruited and enrolled between May 2021 and August 2021 from the 3 affiliated hospitals at Unity Health Toronto. Participants were enrolled if they were an HCW currently providing health care at their respective hospital of employment, aged ≥18 years, and owned a mobile phone (an Android mobile phone with operating system version 6.0 or above or an iPhone with operating system version 11.0 or above).

**Statistical Analysis**

As this was a pilot feasibility trial, we summarized dropout rates, easiness of use, tolerability, acceptability, and utility using counts and percentages. Continuous data were summarized using range, mean and SD, and median and IQR. To assess the
effect of the VR scenario on symptoms of MI, we compared MIOS scores across the 4 time points using a Friedman test. In addition, follow-up MIOS scores were compared with the score at prebrief using Wilcoxon signed rank tests with Bonferroni correction (.05/3=.0167) to adjust for multiple comparisons. As PSS scores were collected only at 2 time points (ie, at prebrief and debrief), a Wilcoxon signed rank test was used to compare the difference in the PSS scores between these 2 time points. A P value of <.05 was considered significant unless otherwise specified. We performed statistical analysis using SAS 9.4 (SAS Institute Inc).

Quantitative Analysis

Stress and MD Analysis

In this feasibility study, we piloted the application of the MIOS to assess MD both acutely and longitudinally. As mentioned in the Study Design section, MIOS was administered during the prebrief, preintervention test, postintervention test, and debrief. Participants were also prompted to complete MIOS on the mobile app in the 8 weeks after the intervention for a longitudinal assessment of MD and MI. All questionnaires used in the mobile app (eg, the MIOS and the PSS) are available in the appendices of the study by Nguyen et al [28].

IPQ Assessment

To objectively assess user experience within the VR scenario, we adopted the IPQ, which is a questionnaire for measuring the sense of presence experienced in a virtual environment [31]. Composed of 14 questions (answered on a 6-point Likert scale), the IPQ has a high reliability (Cronbach \( \alpha = .87 \)) and outputs four items (1 general item, not belonging to a subscale, and 3 subscales): (1) general presence (sense of being there), (2) spatial presence (the sense of being physically present in the virtual environment), (3) involvement (measuring the attention devoted to the virtual environment), and (4) experienced realism (measuring the subjective experience of realism in the virtual environment).

Hereinafter, the 4 outputs will be referred to as IPQ components. More information about the construction and structure of the scale and the IPQ’s reliability analysis is available on the Igroup project consortium website [34,35].

Mobile Data Analysis

After participating in the intervention, participants were instructed by our research staff to download and regularly use our mobile app to answer surveys in the 8-week follow-up. Participants received push notifications on the mobile app 3 times weekly to answer short versions of the scales related to depression (2-item Patient Health Questionnaire), anxiety (2-item Generalized Anxiety Disorder), stress (4-item PSS) MI (4-item MIOS), and loneliness (3-item University of California Los Angeles Loneliness Scale). With the exception of the 3-item University of California Los Angeles Loneliness Scale, participants were also asked to answer the full version of these scales once weekly. Short versions of the scales were used on weekdays to minimize participant burden. The mobile app also had the option of collecting passive data from built-in smartphone sensors (GPS and accelerometer) from participants who provided in-app consent to gather information on distance traveled and activity patterns. Details on the mobile data collection were previously overviewed in the study by Nguyen et al [28]. We used in-app automated survey reminders to promote app use.

Qualitative Analysis

Content Analysis

We performed a content analysis on the data collected from the scenario debriefing conducted immediately after the compound intervention. Qualitative content analysis is a method to interpret meaning from text data and draw conclusions from words, themes, or concepts that occur in the text, in reference to their context, so that research questions can be answered [36]. We used inductive category development by becoming immersed in the data and allowing insights on categories to emerge from the data [37]. The scenario debriefing consisted of a semi-structured interview that allowed participants to answer open-ended questions about their overall experience, followed by a semi-structured debriefing methodology (the interview guide is included in Multimedia Appendix 1). The researchers (BN and AT) who collected the VR data were trained using the Promoting Excellence and Reflective Learning in Simulation (PEARLS) health care debriefing tool [38], a simulation debriefing framework to help learners assess their experience within a safe environment. A flow diagram of the debriefing can be seen in Figure 3. After completion of the intervention, we conducted a postexperiment procedure, which consisted of removing the VR headset from the participant but keeping the physiological sensors attached. In addition, we confirmed with the participant that they were able to continue with the debriefing.

**Figure 3.** Flow diagram of the debriefing. PEARLS: Promoting Excellence and Reflective Learning in Simulation; VR: virtual reality.

During the open-ended feedback part of the debrief, we asked participants to speak freely about their experience with the experiment. We specifically asked the following questions:

1. “What suggestions or feedback would you give to improve the scenarios? Please comment on what can be improved, what can be more realistic, and any deviation from real-life applications.”

2. “Could you share something that you have learned about moral injury today? How might this apply to your clinical practice?”

The research questions we sought to answer with our content analysis from this feedback were as follows:

1. “How can the VR scenario be improved?”

https://games.jmir.org/2024/1/e42813
2. “How accessible and relevant was our intervention?”
We subsequently conducted scenario-based debriefing using the PEARLS [38] methodology, which involved an exploration of the following predetermined topics: participant experience with the technology used, decision-making during the scenario, and emotions elicited during the scenario. The research questions we sought to answer with the content analysis from the scenario-based debriefing were as follows:

1. “What is the overall user experience of participants with the VR technology?”
2. “What were the determining factors for the decisions that participants made in the scenario?”
3. “How did the scenario make the participant feel?”

The PEARLS structure is a well-validated debriefing tool that is typically used to provide introspection on performance for a simulation participant [38]. It has been used extensively in the simulation literature, including a recent user qualitative study with patient-led simulations [39]. A PEARLS debrief integrates 4 main segments: setting the scene, eliciting reactions, description and analysis of the experience, and summary or reflections.

After the debrief, participants were asked to answer a debrief feasibility questionnaire with 3 five-point Likert questions answered on a scale ranging from 1 (strongly disagree) to 5 (strongly agree) about the relevance and utility of the psychoeducational content on MD for real-life situations as well as the ability of the VR scenario to elicit emotions (Multimedia Appendix 1).

User Experience

To assess the user experience within the VR scenario, we evaluated the dropout rate, the feasibility questionnaire, and the qualitative responses provided during the debrief. During the VR scenario, participants had their head and hand movements tracked by the VR headset and controllers, and all movements were mapped into a virtual avatar (Figure 4). To help improve the sense of body ownership (ie, making the users recognize the virtual body as their own) [40], the preintervention test started with a tutorial that had the participants looking at a mirror and moving their head and hands to visualize that their virtual avatar actions reflected their own.

Figure 4. Snapshot of the virtual reality scenario showing the participant’s avatar reflected in a mirror. The blue beam indicates the cursor used to interact with the virtual environment.

Ethics Approval

Ethics approval was obtained from the research ethics board at St. Michael’s Hospital before starting any study activities (21-066).

Results

Participants

Participant Flow

A total of 16 participants were assessed for eligibility; 1 (6%) declined to participate, and therefore 15 (94%) participants were allocated to the intervention. All 15 participants received the intervention. No participants were lost to follow-up, and data from all 15 participants were analyzed. Information on participant flow is presented in Figure 5.
**Baseline Data**

Our sample consisted of 15 HCWs (female participants: n=11, 73%; male participants: n=4, 27%). The participants had a mean age of 32.7 (SD 9.5) years; the male participants had a mean age of 34.3 (SD 4.9) years, whereas the female participants had a mean age of 32.2 (SD 10.9) years. Among the 15 participants, the most common occupations were nursing (n=7, 47%) and medicine (n=3, 20%); other professions included mental health research staff (n=2, 13%), physician assistant (n=1, 7%), educator (n=1, 7%), and graduate student (n=1, 7%). At the time of the experiment, none of the 15 participants had a prior or current COVID-19 infection; however, 4 (27%) had a prior family history of COVID-19 infection. The VR experiments were conducted between May 2021 and August 2021.

**Data Analyzed**

For 15 participants, MIOS, PSS, IPQ, and mobile data were analyzed. The data of 14 participants were analyzed for the content analysis.

**Quantitative Analysis**

**Stress and MD Analysis**

The average MIOS scores for the prebrief, preintervention test, postintervention test, and debrief were 10.4 (SD 9.9), 12.9 (SD 6.9), 12.6 (SD 7.1), and 13.5 (SD 9.1), respectively, with a difference between the debrief and prebrief (between after the intervention and before the intervention) of 3.1 (SD 6.8; Table 1). There was no statistical difference in the MIOS scores at the 5% level when comparing all 4 scores using the Friedman test (Q=4.61; P=.20). Using Bonferroni correction (.05/3=.0167), the results showed no significant difference between the prebrief scores and any follow-up score: preintervention test (P=.30), postintervention test (P=.32), and debrief (P=.11). The MIOS is a new scale that is still under development by the MIOS Consortium and has not yet been established for the assessment of MI [18,41].
Table 1. Wilcoxon signed rank test comparing Moral Injury Outcome Scale follow-up scores at preintervention test, postintervention test, and debrief with the prebrief score (n=15)\(^a\).

<table>
<thead>
<tr>
<th></th>
<th>Values, mean (SD)</th>
<th>Values, median (IQR; range)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prebrief score</td>
<td>10.4 (9.9)</td>
<td>12 (0 to 17; 0 to 28)</td>
<td>N/A(^b)</td>
</tr>
<tr>
<td>Preintervention test score</td>
<td>12.9 (6.9)</td>
<td>13 (6 to 17; 3 to 27)</td>
<td>.30(^c)</td>
</tr>
<tr>
<td>Postintervention test score</td>
<td>12.6 (7.1)</td>
<td>13 (8 to 17; 1 to 28)</td>
<td>.32(^c)</td>
</tr>
<tr>
<td>Debrief score</td>
<td>13.5 (9.1)</td>
<td>14 (5 to 18; 0 to 32)</td>
<td>.11(^c)</td>
</tr>
<tr>
<td>Difference (debrief – prebrief)</td>
<td>3.1 (6.8)</td>
<td>1 (−1 to 7; −8 to 18)</td>
<td>.11(^c)</td>
</tr>
</tbody>
</table>

\(^a\)There was no statistical difference in the Moral Injury Outcome Scale scores at the 5% level when comparing all 4 scores using the Friedman test (\(Q=4.61; P=.20\)).

\(^b\)N/A: not applicable.

\(^c\)Follow-up scores were compared with the preintervention test score using the Wilcoxon signed rank test; Bonferroni correction was used (.05/3=.0167), that is, significance at 1.67% was applied.

PSS scores were only collected at 2 time points: at prebrief and debrief. The average PSS scores during the prebrief and the debrief were 17.3 (SD 7.5) and 19.1 (SD 8.1), respectively, with a postintervention test–preintervention test difference of 1.8 (SD 6.0; Table 2). Similar to the MIOS scores, the prebrief and debrief PSS scores were not statistically different (\(P=.22\)). Tables 1 and 2 summarize the analysis for the MIOS and PSS scores.

Table 2. Wilcoxon signed rank test of the Perceived Stress Scale prebrief and debrief scores (n=15).

<table>
<thead>
<tr>
<th></th>
<th>Values, mean (SD)</th>
<th>Values, median (IQR; range)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prebrief score</td>
<td>17.3 (7.5)</td>
<td>15 (12 to 22; 4 to 33)</td>
<td>N/A(^a)</td>
</tr>
<tr>
<td>Debrief score</td>
<td>19.1 (8.1)</td>
<td>19 (14 to 26; 4 to 33)</td>
<td>N/A</td>
</tr>
<tr>
<td>Difference (debrief – prebrief)</td>
<td>1.8 (6.0)</td>
<td>1 (−1 to 7; −11 to 11)</td>
<td>.22(^b)</td>
</tr>
</tbody>
</table>

\(^a\)N/A: not applicable.

\(^b\)Wilcoxon signed rank test to test no difference in the distribution between the preintervention test and postintervention test scores.

**IPQ Assessment**

On the basis of the data collected from the 15 participants, the VR scenario achieved an above-average degree of overall presence, spatial presence, and involvement, with slightly below-average realism (Table 3 and Figure 6). Considering that the presence component is influenced by the other 3 components, it makes sense that it has a higher variance and SD, which suggests an opportunity to improve the immersion of the VR scenario. The lowest scoring component was realism, with the lowest variance and SD. These findings are corroborated by the qualitative feedback provided during the debrief session, where only 5 (33%) of the 15 participants commented that the environment felt realistic and that they felt immersed in the experience. By contrast, 1 (7%) of the 15 participants stated that they found the environment more immersive than simulation with real people. The participants’ feedback also highlighted other areas for future improvement, particularly regarding the realism component, such as having less restrictive dialogues, making the ICU environment more crowded, improving the voice-over acting features, and having the ICU equipment show patients’ physiological data (eg, heart rate monitor).

Table 3. Igroup Presence Questionnaire data statistics.

<table>
<thead>
<tr>
<th></th>
<th>Values, mean (SD)</th>
<th>Values, median (IQR)</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>General presence</td>
<td>3.80 (1.47)</td>
<td>4.0 (2.0)</td>
<td>2.17</td>
</tr>
<tr>
<td>Spatial presence</td>
<td>3.53 (1.16)</td>
<td>4.0 (1.6)</td>
<td>1.34</td>
</tr>
<tr>
<td>Involvement</td>
<td>3.48 (0.78)</td>
<td>3.5 (1.0)</td>
<td>0.60</td>
</tr>
<tr>
<td>Experienced realism</td>
<td>2.20 (0.67)</td>
<td>2.5 (1.3)</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Mobile Data Analysis
The dropout rates for the study app were very high. Of the 15 participants, 8 (53%) did not perform any survey, whereas 7 (47%) completed at least 1 questionnaire. Instead of answering surveys periodically, only 4 (27%) of the 15 participants had >1 set of survey results. There were not sufficient mobile data to provide informative analysis. In the future, an improved participant engagement strategy is needed to help optimize mobile data collection.

Post Hoc Sample Size Calculation
As an exploratory analysis, we calculated post hoc sample sizes using 2-tailed paired $t$ tests with a significance level of .05 based on the mean differences in the MIOS and PSS scores observed between the respective scores at prebrief and debrief. The common SDs for each score were estimated using the observed larger SD among the 2 scores. The computed correlations between the 2 scores were used in this calculation. The sample sizes required to achieve a power of 80% were 41 and 95 participants based on the observed results for the MIOS and the PSS, respectively (Multimedia Appendix 2).

Qualitative Analysis

Content Analysis
Content analysis was performed for 14 (93%) of the 15 participants because technical issues compromised the speech recording of the 15th participant. Common references to real-life experiences were recognized in the content analysis, with the most frequent themes being the following: the virtual characters’ choices during the experiment were too restrictive (10/14, 71%), feelings of some guilt or shame (8/14, 57%), no feelings of failure or being punished (7/14, 50%), no guilt or shame (6/14, 43%), need of organizational support to deal with the morally challenging situation presented in the experiment (7/14, 50%), numbness (5/14, 36%), and the VR scenario was immersive, real, or engaging (5/14, 36%). Of the 14 participants, 1 (7%; participant 13) provided contradictory responses to feelings of guilt and shame, once saying that they did experience these feelings and once saying that they did not. Furthermore, 2 (14%) of the 14 participants considered the learning experience about MD and MI valuable and useful to their daily practice. A complete summary of the content analysis is provided in Multimedia Appendix 3.

Participants also recommended some specific areas of improvement in the VR scenario; for example, the following suggestions were made by 1 (7%) of the 14 participants: the patient’s vital signs were at a normal range although he was experiencing respiratory failure, the skin color should be consistent with that of the participant (all virtual characters were White), the scenario was unrealistic because other interventions apart from the ICU ventilator should have been portrayed, and photographs of the patient should have been added to better customize the character’s appearance. Finally, 2 (14%) of the 14 participants reported not being able to relate to religious mentions of God in the VR scenario.

User Experience
Although only 3 (20%) of the 15 participants reported prior experience with VR headsets (Multimedia Appendix 4), there were no dropouts during the VR scenario (Figure 5). As we had expected that new VR users could potentially experience nausea or disorientation, participants were reminded multiple times during the prebrief that they could pause or stop the session at any moment. Having said that, of the 15 participants, 14 (93%) did not report any side effects; only 1 (7%) participant reported claustrophobia and slight anxiety at first, but these feelings quickly subsided, and the participant was able to complete the VR scenario without any further side effects or complaints. Finally, all participants agreed that the VR platform and scenario
were easy to navigate (Multimedia Appendix 4). Regarding the
debrief feasibility questionnaire, of the 15 participants, 6 (40%) agreed that they learned about MD and interventions, and 11 (73%) agreed that the knowledge about MD and interventions will help them perform better in real-life events (Multimedia Appendix 5). Although only 8 (53%) of the 15 participants agreed that the VR simulation managed to make them experience the same emotions as they would in a real-life event (Multimedia Appendix 5), during the qualitative debrief, common emotions cited included some guilt, shame, betrayal, and isolation, which are consistent with MD.

Discussion

Principal Findings
In this work, we developed a fully immersive VR scenario to emulate a real experience of a morally distressing situation by HCWs in a simulated ICU setting during the COVID-19 pandemic and assess its acute effects on physiological and psychological parameters as well as longer-term effects on MD. This was followed by an educational video on MD and MI and appropriate mitigation strategies for MD and finally a repetition of the VR scenario in a pretest-posttest design. Because of COVID-19 constraints that resulted in health care settings often being described as a war zone [42], HCWs have been particularly exposed to PMIEs in their work environment during this pandemic [4,7]. However, despite the attention it has gained over the last decade, the concept of MI remains poorly understood. VR is a promising strategy to investigate MI owing to its ability to provide highly controlled virtual environments, personalized and tailored experiences, and full control and monitoring of the participants by the research team. The VR scenario created by the research team involved a complex ethical problem that became unfortunately frequent owing to the strain of the pandemic: prioritizing which patients would receive vital support in the face of the shortage of essential equipment such as ventilators [6]. This situation may be considered morally distressing because participants may witness the transgression of some of their core moral values [12], but it is not considered severe enough to induce MI. To achieve our goals, we performed a thorough quantitative and qualitative analysis of the acceptability, easiness of use, tolerability, and utility of the VR technology using a head-mounted display. To the best of our knowledge, this study is the first to examine the feasibility of using an immersive VR scenario to investigate the psychobiological impacts of a moral stressor on HCWs, as well as to use physiological parameters to predict the severity of stress and symptoms of MD and MI.

The feasibility analysis showed high acceptability of the VR scenario among participants, with no dropouts occurring during the study. Although only one-fifth of the participants (3/15, 20%) had previously used VR, all participants reported that the VR technology was easy to use. Moreover, the tolerability was also high because only 1 (7%) of the 15 participants reported mild transient side effects (claustrophobia); no participants reported nausea, whereas other specific side effects (eg, headache and dizziness) were neither reported by participants nor inquired on by the research team. This finding aligns with the literature showing that the incidence rate of VR-induced side effects is low and ranges between 0.5% and 8% [43], with the most common side effects being nausea, eye strain, and dizziness [43]. Specifically, nausea is reported to have an incidence rate of 5.2% [44], whereas vomiting is considered a rare event with an incidence rate of approximately 2% [45]. These symptoms are defined as cybersickness, a form of motion sickness that may be experienced during immersive VR experiences [44]. In this study, we hypothesize that the lack of nausea and other symptoms of cybersickness may have been due to limited head motion during the VR scenario and to the relatively reduced duration (mean 26.3, SD 2.7 min) of the experiment [46].

Regarding the technical quality of our VR scenario, the IPQ results revealed that the scenario achieved a high degree of general presence and spatial presence, above-average involvement, and slightly below-average realism. Therefore, most of the participants felt immersed and involved in the virtual environment but reported that the experiment was not realistic enough (10/15, 67%). This lack of realism was corroborated by the content analysis, where only approximately one-third of the participants (5/14, 36%) felt that the scenario was immersive, real, or engaging. To improve the experience of realism in virtual hospital environments, future studies could address the limitations pointed out by participants in the qualitative debrief session, such as more realistic ICU settings with equipment displaying patients’ vital parameters and having ethnically diverse virtual characters to be more representative virtual avatars of participants.

Content analysis of the debriefing revealed that feelings of guilt, shame, betrayal, isolation, and failure were commonly reported; these are impairing moral emotions consistent with MD [7,17,47] and might suggest a violation of moral beliefs. This finding suggests that the VR scenario could acutely induce real experiences of mild MD. Interestingly, numbness was mentioned by approximately a third of the participants (5/14, 36%). This feeling could be considered as a consequence of not having real power in relation to a real-world experience; it may also represent an emotional consequence of being exposed to a PMIE [12,18]. We assume that numbness could be related to potential signs of the erosion of moral agency, not in relation to our intervention but to previous real-world experiences of prolonged and repeated stressors and moral stressors. The content analysis revealed that most of the participants (8/14, 57%) reported guilt and shame, which are feelings consistently related to the experience of MD [7,17]. This finding suggests that the moral stressor experienced during the VR scenario could successfully induce some degree of MD. In addition, half of the participants (7/14, 50%) expressed the need for organizational support, an aspect frequently related to MD. Participants suggested that there could be a greater emphasis on organizational dimensions in future simulations, given the expressed need and the alignment with past research on MD [48]. The findings from the content analysis supported our hypothesis that a VR scenario can be successfully used to elicit and discuss real-life experiences and emotions related to MD.

In contrast to the qualitative results, the quantitative analysis did not show significant changes in the MIOS scores between
before and after the experiment. The PSS scores showed the same trend and were not significantly different from baseline, which contradicts our hypothesis that the VR scenario would significantly increase stress levels. Both the MIOS and the PSS focus on symptoms developed over the last month. Although participants were instructed to rate their symptoms at that specific moment, these scales might not have enough sensitivity to capture acute changes in stress and MD symptoms. Alternatively, the changes in MD symptoms may have not been severe enough to induce significant changes in the MIOS scores acutely. Combining our findings from the qualitative and quantitative analyses, we assumed that some degree of MD was experienced by most participants, but we believe that these symptoms were not severe enough to induce MI. This is an important ethical aspect because the VR scenario was designed by specialists in MD and MI to minimize the risk of inducing significant MD in participants.

As MI may develop in the long term, we additionally attempted to use a mobile app to monitor participants for stress and MD and offer psychological support during an 8-week follow-up. Unfortunately, a longitudinal analysis of MD during the follow-up was not possible owing to very low app compliance. It is possible that participants might have developed additional symptoms of MD during follow-up that otherwise could not be captured by our analysis. However, we believe that this is unlikely because no participants requested the psychological support offered in the study. Alternatively, the brief version of the MIOS might not have been sensitive enough to detect slight but important changes in MD that would otherwise be detected by its complete version or by another MD scale. Having said that, this study is a feasibility study with a small sample size, and such an implication is beyond the scope of this work. Finally, the MIOS is still under development; hence, future studies are needed to assess the validity of the MIOS and its brief version.

Mobile app retention proved to be challenging because more than half of the participants (8/15, 53%) did not use the study app, and less than one-third (4/15, 27%) completed at least 1 set of surveys. Our app engagement strategy was based solely on in-app automated reminders and was insufficient to promote participant retention. This finding is supported by recent literature that recommends a combination of different engagement strategies to optimize app use [49,50]. In addition, another possible explanation for the low compliance is that a user-centered design process was not adopted during app development; therefore, the study app may not be particularly targeted to HCWs as the end users [51,52]. Nevertheless, our results are in line with previous research that demonstrates that retention is frequently a great challenge in mobile health studies in both clinical and nonclinical samples [50,53].

Post hoc sample size calculations indicate that a 3-fold and 6-fold sample size is required to reach a power of 80% for the MIOS and the PSS, respectively. With a sample of only 15 participants, our results were underpowered, which may at least in theory explain the nonsignificance of our quantitative findings and the discrepancy between the qualitative and quantitative results. This study was developed during a critical period of the COVID-19 pandemic, with recruitment occurring between May 2021 and August 2021, when contact restrictions were very strict. As the VR intervention required in-person data collection, recruitment proved to be very challenging. Nevertheless, our sample size of 15 participants is appropriate for a preliminary analysis, considering previous VR studies published in PTSD and other mental health disorders [54-57]. Our post hoc sample size calculations may be useful to guide the design of future adequately powered studies using VR in the context of MD and MI.

Limitations

This study has several limitations that must be considered. First, it is a pilot feasibility study with a single arm and a small sample size; thus, the results should be interpreted with due caution. Additional studies with a controlled design are necessary to assess the safety and effectiveness of VR interventions in the assessment of MD and MI. Second, stratification analysis by demographic variables was not possible owing to the reduced sample size; therefore, we were unable to compare symptoms of MI among different subpopulations (eg, nurses and physicians). In addition, our experiments were performed on a purposive sample of only HCWs, thus limiting the generalizability of our findings to other populations. Third, the debriefing methodology used may have also provided a different lens than a traditional qualitative interview or focus group. Fourth, the MIOS and the PSS were used outside of their time frame scope; additional studies should include assessments that focus on acute symptoms of stress and MD. Fifth, a standardized cybersickness scale to assess the side effects within the VR scenario, such as the Virtual Reality Sickness Questionnaire [58], was not used and might have caused underreporting of side effects in this study. Sixth and last, the low app engagement found during the 8-week follow-up hindered an analysis of any potential long-term consequences of the experiment related to MD. Considering that the symptoms of MI may have a late onset, this represents an important limitation to our findings.

Conclusions

The COVID-19 pandemic has challenged the mental health of HCWs, with increased rates of distress, anxiety, and depression being reported. During patient care, ethically difficult situations became common and put frontline HCWs at risk of MD and MI. VR-based interventions are a promising method to address these limitations because they allow for the possibility of developing experiments in safe, personalized, and highly controlled environments. This pilot study investigated the feasibility of using a VR scenario to simulate the experience of a mild morally challenging event for HCWs during the COVID-19 pandemic and to examine participants’ physiological reactions to making morally difficult decisions in a virtual environment. Our results suggest the feasibility of using a VR scenario to simulate real experiences of morally stressful events and elicit genuine responses associated with MD with high acceptability and tolerability. In addition, our VR-based intervention demonstrated utility as a pedagogical tool for teaching possible ways to prevent and mitigate MD. Future studies should be conducted to further validate our findings in a larger sample.
Acknowledgments
This work was funded by Innovation for Defence Excellence and Security (IDEaS), Competitive Projects, Department of National Defence, Canada. The authors are grateful to Dr Deborah Kenny, Ms Kristen Sampson, and the Unity Health Toronto Simulation Program for their contribution and support. The financial support of the Ontario Trillium Scholarship program is gratefully acknowledged by AT.

Data Availability
The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions
The study was conceptualized by VB along with SK, EP, and AD. CWE was involved in the interpretation of results, manuscript write-up, and revision. BN and AT were involved in software design of the virtual reality scenario, data collection, data analysis, and the writing of methods and quantitative results. WS was involved in data collection and logistics. EP was involved in scenario development and qualitative data analysis. LB and DMC provided support with logistics and the debrief component. HJ conducted statistical analysis under the supervision of WL. BK and AD supervised AT for the creation of the virtual reality scenario. SK and VB supervised CWE, BN, and AT on all their tasks.

Conflicts of Interest
VB is supported by an Academic Scholar Award from the Department of Psychiatry, University of Toronto, and has received research support from the Canadian Institutes of Health Research, the Brain & Behavior Research Foundation, Ministry of Health Innovation Funds, the Royal College of Physicians and Surgeons of Canada, the Department of Defence (Canada), and an investigator-initiated trial from Roche Canada. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Postintervention debrief interview guide.
[DOCX File, 11 KB - games_v12i1e42813_app1.docx ]

Multimedia Appendix 2
Post hoc sample size calculations.
[DOCX File, 8 KB - games_v12i1e42813_app2.docx ]

Multimedia Appendix 3
Individual summary of the most common themes in the content analysis of data of 14 participants.
[DOCX File, 12 KB - games_v12i1e42813_app3.docx ]

Multimedia Appendix 4
Results of the virtual reality scenario feasibility questions.
[DOCX File, 9 KB - games_v12i1e42813_app4.docx ]

Multimedia Appendix 5
Scores from the debrief feasibility questionnaire.
[DOCX File, 9 KB - games_v12i1e42813_app5.docx ]

References


31. igroup presence questionnaire (IPQ) overview. igroup. URL: http://www.igroup.org/pq/ipq/index.php [accessed 2022-03-02]
34. igroup presence questionnaire (IPQ) scale construction. igroup. URL: http://www.igroup.org/pq/ipq/construction.php [accessed 2022-03-02]
35. igroup presence questionnaire (IPQ) factor analysis. igroup. URL: http://www.igroup.org/pq/ipq/factor.php [accessed 2022-03-02]


Abbreviations

ECG: electrocardiography
HCW: health care worker
ICU: intensive care unit
IPQ: Igroup Presence Questionnaire
MD: moral distress
MI: moral injury
MIOS: Moral Injury Outcome Scale
PEARLS: Promoting Excellence and Reflective Learning in Simulation
PMIE: potentially morally injurious event
PSS: Perceived Stress Scale
PTSD: posttraumatic stress disorder
VR: virtual reality

©Caroline W Espinola, Binh Nguyen, Andrei Torres, Walter Sim, Alice Rueda, Lindsay Beavers, Douglas M Campbell, Hyejung Jung, Wendy Lou, Bill Kapralos, Elizabeth Peter, Adam Dubrowski, Sridhar Krishnan, Venkat Bhat. Originally published in JMIR Serious Games (https://games.jmir.org), 09.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
A Serious Game to Train Rhythmic Abilities in Children With Dyslexia: Feasibility and Usability Study

Francois Vonthron¹, MSc; Antoine Yuen¹, MSc; Hugues Pellerin², MSc; David Cohen²,³, MD, Prof Dr; Charline Grossard²,³, PhD

¹Poppins, Palaiseau, France
²Service de Psychiatrie de l’Enfant et de l’Adolescent, Groupe Hospitalier Pitie-Salpêtrière, Assistance Publique–Hôpitaux de Paris, Paris, France
³Institut des Systèmes Intelligents et Robotiques (ISIR, CNRS UMR7222), Sorbonne Université, Paris, France

Corresponding Author:
Francois Vonthron, MSc
Poppins
73 rue Leon Bourgeois
Palaiseau, 91120
France
Phone: 33 669515961
Email: francois.vonthron@gmail.com

Abstract

Background: Rhythm perception and production are related to phonological awareness and reading performance, and rhythmic deficits have been reported in dyslexia. In addition, rhythm-based interventions can improve cognitive function, and there is consistent evidence suggesting that they are an efficient tool for training reading skills in dyslexia.

Objective: This paper describes a rhythmic training protocol for children with dyslexia provided through a serious game (SG) called Mila-Learn and the methodology used to test its usability.

Methods: We computed Mila-Learn, an SG that makes training remotely accessible and consistently reproducible and follows an educative agenda using Unity (Unity Technologies). The SG’s development was informed by 2 studies conducted during the French COVID-19 lockdowns. Study 1 was a feasibility study evaluating the autonomous use of Mila-Learn with 2500 children with reading deficits. Data were analyzed from a subsample of 525 children who spontaneously played at least 15 (median 42) games. Study 2, following the same real-life setting as study 1, evaluated the usability of an enhanced version of Mila-Learn over 6 months in a sample of 3337 children. The analysis was carried out in 98 children with available diagnoses.

Results: Benefiting from study 1 feedback, we improved Mila-Learn to enhance motivation and learning by adding specific features, including customization, storylines, humor, and increasing difficulty. Linear mixed models showed that performance improved over time. The scores were better for older children (P < .001), children with attention-deficit/hyperactivity disorder (P < .001), and children with dyslexia (P < .001). Performance improved significantly faster in children with attention-deficit/hyperactivity disorder (β = .06, t3754 = 3.91, P < .001) and slower in children with dyslexia (β = -.06, t3816 = -5.08, P < .001).

Conclusions: Given these encouraging results, future work will focus on the clinical evaluation of Mila-Learn through a large double-blind randomized controlled trial comparing Mila-Learn and a placebo game.

(JMIR Serious Games 2024;12:e42733) doi:10.2196/42733

KEYWORDS
serious game; rhythm; dyslexia; musical abilities; design framework; reading skills; children; digital health intervention; attention-deficit/hyperactivity disorder; ADHD; child development; mobile phone
Introduction

Background
Music training and music-based interventions are becoming increasingly popular for developing brain and cognitive functions in children [1-5]. Building on brain plasticity induced by learning music and the tight link between musical and cognitive skills [6-8], music interventions have been used as training tools in neurodevelopmental disorders (NDDs) such as dyslexia [9-13]. Musical skills, especially when developed in childhood, are associated with enhanced cognitive abilities in various domains, such as attention, processing speed [3], executive functions [14], or speech and language [15-17]. Improvements in cognitive skills induced by musical training have been attributed to structural and functional brain changes in areas that support both music processing and cognition [6-8,18,19].

Recent studies have focused specifically on the relationship between rhythmic skills, such as the capacity to discriminate musical rhythms or synchronize with a beat [20,21], and cognition during development. Tierney and Kraus [22] showed that correlations exist between synchronization with a metronome and attentional and reading skills in typically developing adolescents. In children, rhythm production accuracy is associated with both phonological awareness and reading [23]. Rhythm perception is also related to reading performance [24,25]. Language and music processing may rely on common timing mechanisms that allow for the extraction of temporal information, which is crucial to accurately perceive sequences of events [7,20,26-28]. This hypothesis is supported by neurofunctional evidence as temporal processing involved in music and language recruits partially overlapping neuronal pathways that include the auditory cortex, dorsal premotor cortex, cerebellum, basal ganglia, and thalamus [29,30]. Further evidence of the link between rhythmic skills and cognitive abilities comes from the observation that rhythmic skills are disrupted in NDDs that also affect cognition. Notably, rhythmic deficits have been extensively reported in individuals with dyslexia. Children and adults with dyslexia exhibit inaccurate rhythm perception [25,31] as well as increased variability in motor tapping tasks [32]. These observed rhythmic deficits have given rise to theories (eg, the temporal sampling framework; Goswami [26]) that postulate that poor predictive temporal sampling and coding of events explain reading difficulties in those with dyslexia [26,33].

Building on the importance of rhythmic skills in development, music-based training protocols for children have been developed in recent decades. Studies have shown that children with dyslexia who participate in music-based interventions display better reading and phonological abilities [10-12]. In addition, the effect of music-based programs was extended to typically developing children, who showed significant improvements in speech processing skills and verbal intelligence [14]. However, these encouraging preliminary data have not reached the recommended quality for evidence-based studies owing to methodological limitations such as limited sample size, lack of blind assessment, and potentially inconsistent delivery of interventions [34]. In addition, access to these interventions is still too limited, with inequalities remaining because of significant disparities according to social background and place of residence [35]. For instance, children in poor and remote urban areas, who are more likely to develop an NDD [36], often have less access to care. Furthermore, these traditional music-based interventions usually require in-person instruction, which can be challenging under certain circumstances such as during the COVID-19 pandemic or in areas with limited access to specialized resources. More research is needed to determine whether written language skills can improve in children with dyslexia after training with more accessible and scalable music-based interventions.

To address these limitations, serious games (SGs) designed for educational and training purposes provide a more standardized, scalable, and accessible format for delivering music-based interventions through information and communications technologies. This approach allows for the delivery of the same training to a large sample regardless of geographic location or in-person resource availability. The number of SGs developed for educational and training purposes has increased over the last decade [37], primarily because of the expansion of information and communications technologies such as mobile technologies and telehealth systems. As most households, including those in low-income brackets and rural areas, are now equipped with at least 1 tablet, smartphone, or computer, these SGs can be broadly accessible [35]. Furthermore, a meta-analysis revealed that, across domains, learning is improved with SGs compared with conventional methods [38]. In addition to motivation, several preliminary findings have supported another exciting alternative hypothesis that playing an SG fosters electrical brain activity related to memory, emotions, and concentration [39], providing a possible neuronal explanation for the beneficial effect of SGs. SGs have been used in typically developing populations [40] and in children with NDDs [41,42]. Notably, SGs have been used to deliver rhythm-based training to healthy young adults [43]. Recently, interest in using computer-based interventions to train rhythm skills has been explored in people with dyslexia [44]. One SG named “Jellys” was developed for this purpose in a usability study and showed that children with dyslexia positively engaged with this type of remediation [45]. However, although some studies seem to support the effectiveness of using SGs as a treatment for people with NDDs, the methodological quality of these studies is limited, and further research is needed [46].

Objectives
In this study, our goal was to evaluate the usability of Mila-Learn, an SG aimed at training rhythmic abilities in children with dyslexia. The methodological design of the game was developed in user participatory pilot studies, allowing the children and their families to provide feedback to shape human-machine interactions. We report on 2 studies conducted during the French COVID-19 lockdowns. The first was a feasibility study to assess the children’s engagement through gameplay frequency and collected their feedback. After modifying the SG according to the feedback by adding specific features such as customization, storyline, humor, or increasing difficulty, we report a usability study that addressed the
Methods

Overview

Mila-Learn is an SG that delivers rhythm-based exercise designed for children with dyslexia (called “specific learning disorders in the field of reading” in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition). This SG involves a rich musical universe aiming to lead the child to spontaneously come back and engage with the instrument with their parents. It consists of two main elements: (1) a mobile app that offers rhythmic, sensory-motor, and cognitive tasks in the form of musical activities; and (2) secure servers that allow for data analysis. It enables real-time evaluations to understand children’s difficulties and improve the effectiveness of Mila-Learn. In this section, we first describe the SG from its initial beta version. We then detail the methodology of the 2 exploratory studies conducted during the COVID-19 lockdown.

Mila-Learn Description


The first version of Mila-Learn included five tasks:

1. Dance With Your Hands is an auditory-motor coordination exercise. It involves performing a movement following the tempo of a piece of music. These are pieces with a 4/4 signature that easily allow the child to have rhythmic stability and associate a motor action with the rhythm. This action includes gestures such as clapping, silencing, and raising the arms in the air.
2. Play the Drums is a rhythmic memory game. A drum appears on the child’s screen, and a sequence is played. The child then presses on the drum elements to reproduce the initially played sequence.
3. Rhythmic Vitamins is an exercise in singing and repetition [47]. An initial recorded vocal sequence consisting of syllables and phonemes is played by the software. The child must reproduce it using the rhythm, pronunciation, and pitch of the initial sequence.
4. Following the Tempo requires recognizing and reproducing different rhythmic structures. In this exercise, the child is asked to mark the strong beats of music using the space key on the keyboard.
5. Musical Pitch is an exercise of association between the pitch and its representation. A sound sequence composed of 3 sounds is played (high, medium, and low), and then a visual representation is displayed composed of lines (low, medium, and high). The child has to judge whether the graphic production of the sound is correct.

This prototype version (Figure 1 [48]) was offered to a small group of children end users with dyslexia. We asked them to provide feedback on the design and players. The data collection method was centered on gathering children’s feedback at the end of 15-minute game sessions. A total of 14 children were invited to respond, interact, and provide feedback on the first version of Mila-Learn in the form of a progressive web application [49]. In total, 3 sessions per week over a period of 3 months were conducted.

Figure 1. Landing page of the prototype with 3 different tasks each day [56].
attention-deficit/hyperactivity disorder (ADHD), because of the high co-occurrence between dyslexia and ADHD [50].

**Design Framework**

We integrated new gaming features to increase players’ motivation and engagement. First, we developed a storyline that allowed us to include the different tasks within a larger story. The use of a storyline supported the engagement of a player in the games [51,52]. The storyline was intended to not be complex to prevent the child from losing the main goal of the game [53]. The story in Mila-Learn was designed to correspond to the interests of children aged between 7 and 14 years. The story is inspired by *shonen* manga, which is manga inspired by the cartoon universe. This type of manga is based on a storyline that involves a young hero who starts without knowledge and becomes increasingly powerful during the progression of the story. Some crucial values such as friendship and perseverance are typically present in the story. Most of the time, the first opponent of the hero becomes his friend during the story. In Mila-Learn, the player embodies a little monster who meets another character, a little blue monster named “Blue” who asks for their help—some villagers are held captive by the villain Diabolus, another character. The player has to learn rhythm skills to challenge Diabolus and free the villagers. Over the course of the game, the player discovers that there is a larger villain who holds “Rosa,” the Diabolus’ scooter. At the end, the player must win a large rhythm tournament to finish the game. The story is divided into 12 chapters containing 6 tasks (see the following section). Each task lasts between 1 minute, 20 seconds and 1 minute, 30 seconds, with rare songs playing for 1 minute, 40 seconds to maintain consistency with the music. In this way, we considered the attention capacity of children [53,54].

As recommended in the literature, we created evolving tasks, gradually increasing the level of difficulty in each task and from one task to another [51,53,54]. The tasks must be challenging but accessible. In total, 6 tasks are used in the second version of Mila-Learn. They are introduced progressively to allow the player to practice a task 2 to 3 times before introducing another one. Once all the tasks are known by the player, they increase in difficulty with progression throughout the game. First, within each task, the rhythm displayed corresponds to each beat of a measure. Then, the rhythm changes to correspond to eighth notes (meaning that the rhythm is clapping 3 times in 2 beats) or slows down to be marked only once every 2 beats. Moreover, the songs are played at an increased speed to challenge the player. At higher levels, the marked rhythms can change during the task.

The SG was built to provide clear instructions to the player [51,53]. The instructions are given orally and with visual support, notably by imitating one or more nonplayer characters. Before each task, a quick tutorial allows the player to repeat the movement they have to perform during the game 3 times (ie, clap their hands, touch the screen, and move the tablet). At the beginning of a task, the character played by the user is clearly identified with an arrow. Moreover, for each task, the player is always placed in the same location.

The visual environment is thought to be easily navigated by children. The graphics are pleasant but minimalistic [53,55]. The visuals are thought to be pleasant for children aged between 7 and 14 years and are inspired by the cartoon universe. During the tasks, the background is mostly static, allowing the child to focus on the goal of the task. The characters only move to the rhythm of the music, with repetitive and predictable movements.

We differentiate between short- and long-term goals [51,53,54]. In each task, there is only 1 clear goal (ie, touch the screen to the rhythm) that is clearly differentiated from the long-term goal of a chapter (ie, complete the chapter to challenge Diabolus; Figure 2). Feedback is provided throughout the different tasks using visual cues [51,53,54]. These cues allow the player to know whether they are performing the exercise properly. The feedback for each task is described in the following section. As rewards have been described as a main feature of SGs [51,54], players obtain a reward of 1 to 3 stars at the end of each task depending on their accuracy during the exercise. Personalization has also been described as an important key to enhancing the motivation of the player [51-54]. As in the first version, players have to pick a name for their character at the beginning of the game and modify its color. Finally, we introduced new songs to work on in this version of the game. We added some famous songs known by most children (ie, songs from Disney movies) to increase the motivation of the players. For some tasks such as Fruity Jump, Karate Fruit, and Sing Lab, the predetermined structures of these songs did not make their use possible. We specifically composed songs to fit with the requirements of these tasks.
Description of the Tasks

All the tasks (Figure 3) were designed to work on rhythm, which was the main and explicit goal of each task. However, each task requires the mobilization of other skills such as attention, inhibition, working memory, and motor skills, which are also often impaired in children with dyslexia [50,56].

Follow me aims to introduce rhythm to the player. The child first sees a little monster clapping hands to the rhythm of a song and then has to touch the screen to the rhythm by imitating the monster. Then, the character stops clapping, and the player has to keep going alone without the support of the monster. This task allows the player to work on maintaining regularity in rhythm but also sustained attention.

In Clap Trap, 2 characters and the player appear on the screen. The first 2 characters clap one after another, giving a tempo to the player, who has to complete the sequence by clapping their hands to the rhythm at the right time. The first character claps on the first beat of a 4-time measure of the song played. The second character claps on the second beat, and the player has to clap their hands on the third beat. The microphone records the child’s clap. In this task, the child has to anticipate and adapt to the rhythm. It was designed to train inhibition skills as the child has to wait until the right moment to clap their hands.

In River Splash, the player is placed behind 2 other characters who run next to the water and sometimes have to jump across the river to the rhythm. The first character jumps on the first beat of a 4-time measure of a song. The second character jumps on the second beat, and the player jumps on the third beat. The player has to shake the tablet quickly to jump. In addition to rhythm perception, this task was designed to train inhibition skills similarly to the Clap Trap task.

In Sing Lab, the first character produces a sequence of phonemes or syllables at a particular tempo. The player has to reproduce this sequence with particular attention to the pattern and duration of the phonemes. Phonemes or syllables pass across the top of the screen, visually represented by gauges that the player has to fill. If the child sings at the right time, the gauge starts filling. When the duration of the note is complete, the gauge changes color from white to green. In this task, the phonological loop is involved in correctly memorizing the sequence. We used specific music constructed for this task that allowed us to add phonemes or syllables to sing at particular moments and for as long as we wanted.

In Fruity Jump, a character reproduces a rhythmic sequence. The player has to memorize this sequence and then reproduce it correctly by tapping the screen at a good tempo. The tempo is visually indicated by fruits falling from a tree. If the player claps at the right time, the character jumps and hits a fruit with its head to throw it to another tree. If the player misses the fruit, it crashes on the ground. If the player jumps at another time (ie, during the demonstration), their character jumps, and nothing special happens. Working memory and intermodality are particularly engaged in this task.

Finally, in Karate Fruits, the player has to hit fruits that appear based on the rhythm. To hit the fruit, the player has to put the tablet on the floor and extend their arms above it. Each time the camera detects the arm, the player’s character punches. If the player punches at the right time, the fruit explodes, and a smiley face appears. If the player misses the fruit, the fruit goes off the screen, and a smiley face with an annoyed head appears. If the player punches at another time, the character punches, and nothing else happens.
Figure 3. Examples of screens in Mila-Learn: (A) Clap Trap, (B) River Splash, (C) Sing Lab, (D) Fruity Jumps, (E) Karate Fruits, and (F) Follow Me.

Scoring Player Performance

Scoring of player performance is based on rhythmic synchronization through multiple modalities of interaction (sometimes in combination) as rhythmic synchronization is a requirement for all games. Player responses are captured through accelerometers, microphones, webcams, and pressure-sensitive screens, as shown in Table 1.

By assessing the audiomotor synchronization of the child with the rhythmic instruction, we define (1) a time T that corresponds to the exact moment when the player’s input is expected (regardless of the interaction mode) and (2) tolerance thresholds (t_{Perfect}, t_{Good}, t_{Correct}).

The different intervals allow for judging the quality of the answer with 4 levels of acceptance. An input is considered acceptable when it is in the interval $[T - t_{Correct}, T + t_{Correct}]$. 
not acceptable otherwise. An input of better quality, either in the interval $[T - t_{\text{Good}}; T + t_{\text{Good}}]$ or in the interval $[T - t_{\text{Perfect}}; T + t_{\text{Perfect}}]$, results in different visual and audio feedback for the child.

In the second improved version, which was a modified version based on the first pilot study, a simplified calculation was performed by considering the ratio of acceptable inputs to total inputs as the main measure. This final score is presented to the child in the form of stars depending on their performance: no stars if the child has an average of <50%, 1 star if ≥50% of inputs are acceptable, 2 stars for ≥75%, and 3 stars for ≥90%. In addition, this architecture allows for the storage of all the child’s inputs for retro-analysis purposes.

### Table 1. Players’ recorded responses and game parameters in the second version of Mila-Learn.

<table>
<thead>
<tr>
<th>Task</th>
<th>Type</th>
<th>Interaction</th>
<th>Capture technology</th>
<th>Songs</th>
<th>Tolerance threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow Me</td>
<td>Continuous tapping</td>
<td>Tapping</td>
<td>Contact pressure</td>
<td>Commercial</td>
<td>• $t_{\text{Perfect}}$: 0.1 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Good}}$: 0.15 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Correct}}$: 0.25 s before or after the beat</td>
</tr>
<tr>
<td>Clap Trap</td>
<td>Last beat</td>
<td>Clapping hands</td>
<td>Microphone</td>
<td>Commercial</td>
<td>• $t_{\text{Perfect}}$: 0.1 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Good}}$: 0.15 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Correct}}$: 0.25 s before or after the beat</td>
</tr>
<tr>
<td>River Splash</td>
<td>Last beat</td>
<td>Shaking tablet</td>
<td>Accelerometer</td>
<td>Commercial</td>
<td>• $t_{\text{Perfect}}$: 0.1 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Good}}$: 0.15 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Correct}}$: 0.25 s before or after the beat</td>
</tr>
<tr>
<td>Sing Lab</td>
<td>Call and response</td>
<td>Singing</td>
<td>Microphone</td>
<td>Commercial+built in-house</td>
<td>• $t_{\text{Pattern}}$: 0.15 s before or after the beat and up to 30% of the note duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Song duration: the note must be sung at least 60% of the time</td>
</tr>
<tr>
<td>Fruity Jump</td>
<td>Call and response</td>
<td>Tapping</td>
<td>Contact pressure</td>
<td>Built in-house</td>
<td>• $t_{\text{Perfect}}$: 0.2 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Good}}$: 0.25 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Correct}}$: 0.3 s before or after the beat</td>
</tr>
<tr>
<td>Karate Fruits</td>
<td>Last beat</td>
<td>Punching</td>
<td>Webcam</td>
<td>Built in-house</td>
<td>• $t_{\text{Perfect}}$: 0.08 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Good}}$: 0.14 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• $t_{\text{Correct}}$: 0.3 s before or after the beat</td>
</tr>
</tbody>
</table>

$a$Name of the task.

**Feasibility Study**

We conducted a feasibility study to evaluate whether children with NDDs involving reading deficits could use Mila-Learn autonomously at home. Our main objective in assessing Mila-Learn’s autonomous use was to monitor both the time users spent on the SG and their accuracy in each game played. In the context of the unprecedented health crisis caused by COVID-19, participants were recruited by the French Federation for Learning Disorders (FFDys), a national association that aggregates all regional associations of people with learning disabilities. The FFDys communicated to its members the possibility of testing an app and managed the information and consent of participants. Families were informed that Mila-Learn was an SG for performing rhythmic tasks at home and that we believed this practice might be beneficial for learning to read. In total, 2500 children downloaded Mila-Learn. The analyses were conducted on a subsample of 21% (525/2500) of these children, who spontaneously played at least 15 games. To improve the usability of Mila-Learn, we also asked users (both children and families) to provide feedback on the games and information on the children’s impairments. This information was provided freely and was not compulsory to obtain Mila-Learn. In addition, we systematically collected through phone interviews all the problems that the children and their families encountered regarding the computing and web performance of the SG. Finally, we conducted a phone survey of 200 users, which is provided in Multimedia Appendix 1 [56]. The questions asked were designed to gain insights into the families’ perceptions of the benefits of the tool, the improvements and difficulties of use they encountered, and their desire to continue using the game in the future; in addition, room was left for unstructured testimony. The data analysis for the feasibility study was limited to descriptive statistics.

**Usability Study**

This usability study was considered a continuation of study 1 and was conducted under the same ethical rules. In the usability study carried out in a real-life setting over 6 months, our primary focus was 2-fold following modifications to Mila-Learn based on study 1 feedback: first, to ensure that the computational architecture and final version of Mila-Learn were free of computer bugs and, second, to track player progress using Mila-Learn’s scoring system over an extended duration. As part of the second lockdown because of the COVID-19 crisis, the final version of Mila-Learn was made available again starting...
on October 10, 2020, on National Learning Disabilities Day. Benefiting from the large amount of feedback received during the first lockdown, very few technical problems occurred, resulting in a game with much better fluidity that provided higher-quality data. A total of 3337 children had access to Mila-Learn for a total of 84,682 games that were played. As in study 1, at the time of registration, the patients’ families were given the opportunity to complete the profile of the children, including information such as the children’s diagnoses. A total of 304 diagnoses were reported by the parents. Finally, the children and their families had the option of linking the game character to the reported clinical profile. This option was exercised by 2.94% (98/3337) of the children, for whom we had both their reported diagnosis and game performance over time. These 2.94% (98/3337) of the children completed 3922 games.

To assess how children performed with Mila-Learn, we defined and computed the following variables:

1. “Time” is an incremental value representing the number of levels played by a player since the beginning of the experiment. Time is 1 at the beginning of the experiment and represents the total number of levels played by the player at the end of the experiment.

2. “Delta_tap” is the delay between the date of the played input (as defined in Table 1) and the date of the expected input.

3. “Threshold” is a delay defined for each game that was used to construct the performance score.

4. “Performance score” is a variable bounded between 0 and 100 that was created to quantify performance from delta_tap and normalize performance across games. We used the following formulas: performance score = \((-100 / \text{threshold}) \times \text{abs}(\text{delta}_t\text{ap}) + 100\) for \(\text{abs}(\text{delta}_t\text{ap})\leq \text{threshold}\) and \(\text{performance score}=0\) for \(\text{abs}(\text{delta}_t\text{ap})>\text{threshold}\).

We conducted several linear mixed models. To assess children’s progress over time, we tested whether players improved their performance through the progression across the games using a linear mixed model with the following formula: performance score = time + (1|PlayerID/GameID/LevelName).

To assess whether a declared diagnosis was associated with the average performance of the children, we also conducted a linear mixed model using the following formula: performance score = age + dyscalculia + dysgraphia + dyslexia + dysphasia + ADHD + ExecutiveFunction impairment + (1|GameID/LevelName).

Finally, we also tested whether progress over time was moderated by a declared diagnosis using the following formula: performance score = time + diagnosis + time × diagnosis + (1|GameID/LevelName).

Ethical Considerations

Under French legislation, we did not need the approval of a Comité de Protection des Personnes (Committee for the Protection of Persons). However, as the pilot study was conducted in line with the creation of large databases, we obtained the approval of the Commission Nationale de l’Informatique et des Libertés (National Commission for Informatics and Freedoms) under number 2222283.

Results

Feasibility Study

Between April 2020 and June 2020, a total of 2500 children had access to Mila-Learn. Families reported the child’s diagnosis in 60% (1500/2500) of cases. As children were recruited through the FFDys, they were diagnosed with an NDD in almost all cases, but only 23% (575/2500) were declared as having dyslexia. The other children had developmental coordination disorders (dyspraxia), dyscalculia, and communication disorders of oral language (dysphasia). In addition, 18% (450/2500) declared a diagnosis of ADHD.

Data regarding the use of Mila-Learn by each user were recorded as time spent on the SG and accuracy in each game played. The average use was 3.5 sessions per week. To ensure the significance of the data, we only kept the data of players who participated over a sufficient period (>15 games). Duration was expressed as the number of games played. We considered the number of games played inside the SG over the number of played sessions as the number of games played in 1 session could vary widely. In total, 21% (525/2500) of players aged 6 to 14 years played at least 15 games, with an overall mean of 54.77 and a median of 42 games played. The average number of games played was similar across ages (no main effect of age). No effect of age was found on the mean score. In addition, no floor or ceiling effects were observed (Multimedia Appendix 2).

It should be noted that several technical issues occurred during the first 2 weeks owing to the wide variety of tablet operating systems. This situation resulted in the deployment of corrective patches, but owing to the correction delay, it may have differentially altered one child’s experience relative to another’s. To improve the user experience, phone calls were systematically conducted to interview families, determine potential problem areas, and gather feedback for improvement. Parents consistently highlighted the recreational side of the game and its impact on the children’s self-confidence. A survey of 200 users, provided in Multimedia Appendix 1, also indicated that 96% (192/200) wished to continue using Mila-Learn after the COVID-19 pandemic. However, they also provided significant feedback (164/200, 82%) to improve the game. Multimedia Appendix 3 [57] presents the most significant feedback with a frequency of ≥10 occurrences. We classified it according to the criteria by Morville [56], which distinguish 7 dimensions: usefulness, usability, findability (the ease of locating a feature or a piece of context), credibility, accessibility design, attractiveness, and value [48]. Usability was questioned in several comments, such as “the detection of movements should be improved,” “sound detection needs to be improved,” and “the game needs to be better adapted to the child’s difficulty profile.” Accessibility was also questioned as several parents indicated that “the writing could benefit from being larger and the display of dialogues slower.”
Mila-Learn Description Adjustments Following Study 1

Design Framework

On the basis of the feedback obtained during study 1, we made several modifications to Mila-Learn. To improve accessibility, the first modification was to offer the player the choice between several fonts, including OpenDyslexic. This choice is reversible throughout the game. We also improved sound and movement detection. A second significant choice was to distinguish the children’s pathways according to their predominant disorders to facilitate their entry into the game and usability. For example, a player who indicated that they had dyspraxia at the time of registration was offered more moderate motor exercises (ie, Sing Lab), allowing them to enter the adventure before training on River Splash or Karate Fruits that are more challenging in terms of motor abilities. In contrast, a child with dyslexia could be offered River Splash from the beginning, with Sing Lab exercises being offered only afterward as Sing Lab involves the phonological loop.

To increase motivation and interest in the game, we provided new possibilities of personalization for the character: the player could choose the gender of the avatar, their color, and the color of the hat. We then increased the storyline with the help of a screenwriter. This modification improved the consistency of the story and made it more inclusive by adding new characters that could help the player during the game. A new companion named “Mila” appeared, who is a fairy representing the planet “Mila” where the story takes place. These modifications also influenced (1) the dialogues, which were shortened with the language adapted to children; and (2) the appearance of the notion of “rhythmic,” which was introduced as a martial art based on rhythm to clarify the main goal of the game during the adventure.

Through this expansion, we created 6 new chapters. We maintained the same concept as the preceding version and gradually increased the level of difficulty during the progression of the game by increasing the speed of the rhythm and varying the type of rhythm clapped (ie, clapping notes, then eighth notes). We also created daily missions. These 4 daily tasks allowed the child to revisit games on which they had practiced in the past and where they encountered difficulties. This allowed us to directly address the tendency to forget what has been learned and allowed for longer practice with Mila-Learn.

Description of the Tasks

Finally, we made structural modifications to the proposed tasks to ensure the game’s fluidity and improve motor interactions. First, we changed the way children had to answer during Clap Trap. Instead of clapping both hands, which was recorded using the microphone, we changed the child’s interaction with the SG to synchronously tapping both hands on the screen (and, therefore, we used a touch recording). Second, Follow me was extensively modified to be more understandable and involve the child more on a motor level. The interaction was changed from a passive mode (one contact pressure) to a more active hand clapping measured using the microphone. The child did not perform the task all at once but interacted with the character, who gave them instructions that the child reproduced on the principle of call and response. Specific music was created for the game. As a consequence, Follow me was renamed Clap Hero. Finally, we modified the way children had to answer during Fruity Jump—children’s interaction with Mila-Learn changed from tapping to shaking the tablet, which was measured using an accelerometer. Table 2 summarizes the changes made in the final version of Mila-Learn.
Table 2. Players’ recorded responses and game parameters in the final version of Mila-Learn.

<table>
<thead>
<tr>
<th>Task</th>
<th>Type</th>
<th>Interaction</th>
<th>Capture technology</th>
<th>Songs</th>
<th>Tolerance threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clap Hero</td>
<td>Call and response a</td>
<td>Clapping hands</td>
<td>Microphone</td>
<td>Customized</td>
<td>• Perfect: 0.1 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Good: 0.15 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Correct: 0.25 s before or after the beat</td>
</tr>
<tr>
<td>Clap Trap</td>
<td>Last beat</td>
<td>Tapping on the left and right side of the screen</td>
<td>Touch</td>
<td>Commercial</td>
<td>• Perfect: 0.1 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Good: 0.15 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Correct: 0.25 s before or after the beat</td>
</tr>
<tr>
<td>River Splash</td>
<td>Last beat</td>
<td>Shaking tablet</td>
<td>Accelerometer</td>
<td>Commercial</td>
<td>• Perfect: 0.1 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Good: 0.15 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Correct: 0.25 s before or after the beat</td>
</tr>
<tr>
<td>Sing Lab</td>
<td>Call and response</td>
<td>Singing</td>
<td>Microphone</td>
<td>Customized</td>
<td>• Pattern: 0.15 s before or after the beat and up to 30% of the note duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Song duration: the note must be sung at least 60% of the time</td>
</tr>
<tr>
<td>Fruity Jump</td>
<td>Call and response</td>
<td>Shaking tablet</td>
<td>Accelerometer</td>
<td>Customized</td>
<td>• Perfect: 0.2 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Good: 0.25 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Correct: 0.3 s before or after the beat</td>
</tr>
<tr>
<td>Karate Fruits</td>
<td>Last beat</td>
<td>Punching</td>
<td>Webcam</td>
<td>Customized</td>
<td>• Perfect: 0.08 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Good: 0.14 s before or after the beat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Correct: 0.3 s before or after the beat</td>
</tr>
</tbody>
</table>

aItalics indicate game and functional changes that were introduced compared with the Mila-Learn second version summarized in Table 1.

Usability Study

This usability study focused on a sample of 98 children (mean age 9.05; SE 2.4 years), and we had both their reported diagnoses and game performance over time. These 98 children completed 3922 games. The linear mixed models yielded the following significant results. First, we found that the performance of the children significantly improved over time ($\beta=.02; t_{3268}=2.68; P=.007$). That is, there was an increase in the performance score by an average of 5 points after 250 levels were played.

Second, we explored whether declared diagnosis and age influenced the average performance of the children. Table 3 summarizes the results. We found that older children performed better than younger children. One year of age increased the normalized performance score by 1.08 points (meaning 1.1% of the maximal range). In addition, children with dyslexia and ADHD performed significantly better than those with other diagnoses (performance improved significantly faster in children with ADHD, $\beta=.06; t_{3754}=3.91; P<.001$) and that children with dyslexia ($\beta=-.06; t_{3816}=-5.08; P<.001$) and executive dysfunction ($\beta=-.03; t_{3805}=-2.09; P=.04$) improved less over time than those with other diagnoses. Having dyslexia increased the normalized performance score by 2.81 points (meaning 2.8% of the maximal range) compared with children without dyslexia, whereas having ADHD increased the normalized performance score by 4.16 points (meaning 4.2% of the maximal range) compared with children without ADHD. In contrast, children with executive function impairment and dysgraphia performed significantly worse than those with other diagnoses. Having dysgraphia decreased the normalized performance score by 2.06 points (meaning 2.1% of the maximal range) compared with children without dysgraphia, whereas having executive function impairment decreased the normalized performance score by 3.26 points (meaning 3.3% of the maximal range) compared with children without executive function impairment.

Finally, we also tested whether progress over time statistically interacted with the declared diagnosis. We found that children with ADHD progressed faster over time than those with other diagnoses ($\beta=.06; t_{3754}=3.91; P<.001$) and that children with dyslexia ($\beta=-.06; t_{3816}=-5.08; P<.001$) and executive dysfunction ($\beta=-.03; t_{3805}=-2.09; P=.04$) improved less over time than those with other diagnoses. We found no significant interaction between time and a diagnosis of dysphasia ($\beta=-.01; t_{3816}=-0.68; P=.50$), dyscalculia ($\beta=.05; t_{3787}=1.46; P=.14$), or dysgraphia ($\beta=.00; t_{3816}=0.1; P=.92$).
Discussion

Principal Findings

The literature on SGs, especially when designed for a specific medical condition, is limited when it focuses on game design methodology or formal clinical validation [42]. In this paper, we described the process and empirical studies to address this issue for Mila-Learn, an SG based on rhythmic training for children with dyslexia. To do so, we placed the patient’s experience at the center of the game construction while iterating with clinicians involved in treating children with dyslexia. In this paper, we described the different developmental phases that helped us design the game. We first constructed an initial prototype based on a literature review and with the help of clinicians specializing in learning disorders. Then, based on a first round of feedback from users and comments from professionals, we developed a first version of Mila-Learn for tablets.

In this version, we greatly improved the users’ experience with the game by adding new gaming features to increase the motivation and engagement of players. We offered more possibilities for customization, created a storyline, and introduced humorous and friendly characters to align with children’s interests [51,52]. Moreover, we adapted the difficulty of the game to enhance the learning possibilities of children by working on graphism and the instructions given to the children and by creating evolving tasks that gradually increased the level of difficulty [51,53,54]. With this second version, we adopted a user participatory design by inviting children, families, and professionals to test this version and send us feedback about their experience (feasibility study). User participatory design is a method that is currently gaining attention. Contrary to user-centered designs, which create games for a user, participatory design aims to construct the game with the users by collecting their experience and advice and then including them in the game [58]. It has been shown that participatory design promotes engagement of the user [52]. Indeed, collecting feedback both from families and children and from professionals is essential as professionals and families and children focus on different aspects of an SG and do not place the same importance on each feature [52]. We believe that this participative process helped us develop an SG that improved the experience within the game and the interest of families and children in Mila-Learn.

Regarding computational aspects, we also collected feedback that helped us resolve bugs and record the time spent on the game and the player’s accuracy in each game. These features allowed us to follow children’s interest in the game and their progression over time and demonstrate that progression occurred with Mila-Learn and was associated with age. Study 1 confirmed that children could engage with Mila-Learn for a rather long period and play at home without the need for an extra supporting person, suggesting that Mila-Learn was sufficiently motivating and adapted to this population. Children and their families appeared to be highly satisfied with the game.

Finally, following a third round of feedback from parents, children, and professionals, we developed a final version of Mila-Learn to improve accessibility and motivation for the player. We made structural modifications to the proposed tasks to ensure the fluidity of the game and improve motor interactions. We resolved most of the technical problems, which allowed us to conduct a real-life usability study of the Mila-Learn game during the second lockdown.

Comparison With Prior Work

In the usability study, we observed that children significantly improved their scores on the 6 games included in Mila-Learn. Although we cannot conclude that the rhythm abilities of the children improved based only on these results, we believe that the children learned how to use Mila-Learn and that they were increasingly accurate in responding to each game. However, the effect size was small, although it may have been underestimated as the difficulty in the games increased, which could have masked the children’s progression. In addition, based on the diagnosis declared by the children’s parents, we performed exploratory analyses to assess whether improvements over time were associated with the declared diagnoses. Linear mixed models showed that children’s performance significantly increased over time, that scores were better for children with ADHD and dyslexia, and that performance improved significantly faster for children with ADHD and slower for children with dyslexia.

Regarding the average performance of children according to diagnosis, the results were very encouraging if we consider the

### Table 3. Average scores according to diagnosis during study 3 with the final version of Mila-Learn.

| Estimate (SE) | t test (df) | Pr(>|t|) |
|--------------|------------|---------|
| Intercept    | 37.83 (4.60) | 8.23 (6.04) | <.001 |
| Age          | 1.08 (0.15)  | 7.02 (2082.28) | <.001 |
| Dyscalculia (yes) | -0.43 (1.03) | -0.42 (2081.02) | .68 |
| Dysgraphia (yes) | -2.06 (0.79) | -2.60 (2084.61) | .009 |
| Dyslexia (yes) | 2.81 (0.70)  | 4.03 (2087.52) | <.001 |
| Dysphasia (yes) | -0.51 (0.93) | -0.54 (2086.54) | .59 |
| ADHD (yes)   | 4.16 (0.61)  | 6.79 (2090.71) | <.001 |
| Executive function impairment (yes) | -3.26 (1.36) | -2.40 (2086.78) | .02 |

aADHD: attention-deficit/hyperactivity disorder.
relationship between reading impairments and diagnosis. On the basis of the literature, we expected reading impairments to be associated with dyslexia, attention deficit, and specific oral language impairment (dysphasia) [59,60]. In addition, we expected that severity would negatively influence performance. This is usually the case when children have dysphasia [61], executive function impairment [62], or multidimensional impairments [63]. The results were in line with these expectations. Children with dyslexia and ADHD showed a significantly better performance over time, whereas children with dysphasia, executive function impairment, and dysgraphia showed a worse performance. Of note, children with dysgraphia often have motor coordination disorders [64]. Finally, dyscalculia had no influence on Mila-Learn performance. In summary, the predictions according to diagnosis were in line with the hypothesis that Mila-Learn may improve performance in children with reading impairments. The fact that dyscalculia showed no specific effects and that the diagnoses associated with the highest severity (dysphasia and executive function impairment) showed less improvement followed our hypothesis [61-63,65]. We speculated that dysgraphia was associated with multidimensional impairments, including motor coordination disorder. This interpretation is based on the fact that recruitment for the study was based only on user reading impairments.

Regarding the average performance of children according to diagnosis over time (ie, the statistical interaction), the fact that performance improved significantly faster for children with ADHD and slower for children with dyslexia is not surprising as the perception of rhythm is impaired in children with dyslexia. In contrast, children with ADHD may have impairments in reading abilities but do not have specific deficits in rhythm and speech perception [57,66].

Strengths and Limitations

The exploratory studies presented in this paper have some limitations despite the promising results. On the one hand, some aspects of the game need to be improved.

First, we currently consider a “standard” latency of 20 ms, which corresponds to the estimated delay between the child’s real input and the input processed by the operating system. In reality, each tablet may have unique differences. In the next version of Mila-Learn, we need to consider this unique latency to get as close as possible to the real performance of children. This adjustment might lead to more accurate measurements of children’s interactions and, potentially, more tailored game experiences.

Second, the game gradually increases in difficulty with the progression of the player within the game. We integrated some specific pathways as a function of the difficulties that the children declared before starting the game (ie, children with motor difficulties do not start with games that require a high level of motor skills). However, the progression is predetermined and does not take into account the results of the player. In the next version of Mila-Learn, the difficulty of each game will automatically adapt based on the child’s performance in the previous games using a specific algorithm [53], allowing for much better stimulation. By doing so, the game could offer a more individualized experience, potentially leading to more sustained engagement and greater benefits for the children.

Third, the age range of 7 to 14 years is wide as children’s interests can vary greatly during these years. In a future version of Mila-Learn, the graphics and music will be adapted to the age specified by the child so that the game will be more suitable for their age. This may enhance the game’s appeal to players across the entire age range, fostering increased engagement and learning.

However, our study was only exploratory in nature. First, even if the lockdown gave us the opportunity to have a large sample for exploratory studies, diagnoses were not clinically grounded and were only declared by the children’s parents. Therefore, caution should be the rule when interpreting predictive models.

Second, there was no predefined design for the studies as the training was spontaneous and included no comparison with alternative treatment proposals. Therefore, the clinical interest of Mila-Learn for dyslexia cannot be established based on the results of the 2 exploratory studies presented in this paper.

Future Directions

To address the clinical relevance of Mila-Learn in relation to dyslexia, the next step will be to evaluate the effects of Mila-Learn in the context of a randomized controlled trial. Children with dyslexia based on objective clinical assessments will be randomized to Mila-Learn sessions or placebo game sessions that take place in the same universe but with different tasks. We will assess the evolution of reading skills from before to after training with the hope of greater improvements with Mila-Learn. On the basis of the exploratory studies, we calculated the number of patients per group that would ensure a statistical power of at least 85% for an effect size equal to 0.5 (moderate) when the changes in the experimental and control groups were compared. This calculation indicated that each group should have at least 73 children (ie, 146 children in total). This study started in September 2021 (Comité de Protection des Personnes registration 2021-A01709-32; ClinicalTrials.gov Identifier: NCT05154721).

Conclusions

We presented how we constructed Mila-Learn, an SG based on rhythm activities, to improve reading skills in children with dyslexia. We developed several versions of the game considering the literature, professionals’ experiences, and users’ feedback. We also conducted a usability and a feasibility study to evaluate each version of Mila-Learn. The results indicated that Mila-Learn was attractive and sustained the players’ motivation and engagement for several months. Moreover, children were able to learn how to use the game, and their performance in the games improved with training. Future research will include (1) adapting to the latency of the electronic devices, (2) automatically adapting the games based on the player’s performance, and (3) conducting a large randomized controlled trial to evaluate the impact of Mila-Learn on reading skills.
Acknowledgments
The authors would like to thank Valentin Begel for his counsel in the design of this project, the Le Kremlin-Bicêtre hospitals for helping them develop the first version of Mila-Learn, and the École Polytechnique for its financial support. The authors would also like to thank all the families, children, and professionals for their feedback and encouragement. Special thanks to the French Federation for Learning Disorders. This work was supported by the École Polytechnique (grant “Prix Gérondeau 2018”). The sponsors of the aforementioned study funding source were not involved in the study, writing of the report, or decision to submit the paper for publication.

Data Availability
The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions
FV conceptualized the study with DC and CG, managed data curation, acquired funding, undertook the investigation, and contributed to software development alongside AY. AY participated in data curation, formal analysis, and software development. HP was responsible for formal analysis and visualization. DC further contributed to conceptualization and was involved in methodology, supervision, and the review and editing process. CG also took part in conceptualization, formal analysis, writing of the original draft, investigation, supervision, and the review and editing process.

Conflicts of Interest
FV reports a relationship with bMotion Technologies that includes equity or stocks. AY reports a relationship with bMotion Technologies that includes employment and equity or stocks.

Multimedia Appendix 1
User survey on Mila-Learn.
[DOCX File, 26 KB - games_v12i1e42733_app1.docx ]

Multimedia Appendix 2
Average scores from all games between April 2020 and June 2020 according to the children’s age.
[PNG File, 22 KB - games_v12i1e42733_app2.png ]

Multimedia Appendix 3
Feedback classification based on the criteria by Morville [55].
[DOCX File, 27 KB - games_v12i1e42733_app3.docx ]

References


33. Tallal P. Improving language and literacy is a matter of time. Nat Rev Neurosci 2004 Sep;5(9):721-728. [doi: 10.1038/nrn1499] [Medline: 15322530]


47. Bolduc J, Rondeau J. Rythmos les apprentissages !. Langage et pratiques 2015;56:12-22 [FREE Full text]


Abbreviations

ADHD: attention-deficit/hyperactivity disorder
FFDys: French Federation for Learning Disorders
NDD: neurodevelopmental disorder
SG: serious game

©Francois Vonthron, Antoine Yuen, Hugues Pellerin, David Cohen, Charline Grossard. Originally published in JMIR Serious Games (https://games.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
Original Paper

A Novel Casual Video Game With Simple Mental Health and Well-Being Concepts (Match Emoji): Mixed Methods Feasibility Study

Russell Pine¹, PhD; James Mbinta¹, PhD; Lisa Te Morenga², PhD; Theresa Fleming¹, PhD

¹School of Health, Victoria University of Wellington, Wellington, New Zealand
²Research Centre for Hauora and Health, Massey University, Wellington, New Zealand

Corresponding Author:
Russell Pine, PhD
School of Health, Victoria University of Wellington
Kelburn Parade, Kelburn
Wellington, 6012
New Zealand
Phone: 64 4 472 1000
Email: russell.pine@vuw.ac.nz

Abstract

Background: Adolescence is a crucial phase for early intervention and prevention of mental health problems. Casual video games are popular and have promise as a novel mechanism for reaching young people, but this potential has seldom been explored.

Objective: This study aimed to explore the acceptability, feasibility, and possible indicators of therapeutic changes after playing a purpose-built novel casual video game (Match Emoji) with simple mental health and well-being content among young adolescents.

Methods: We conducted a single-arm, nonrandomized trial of Match Emoji with 12- to 14-year-old school students (N=45; 26 [57%] New Zealand European, 12 [26%] Māori; 7 [15%] Asian or Pacific; 27 [60%] boys, 3 [6%] non-binary). Participants were invited to play Match Emoji for 15 minutes, 2-3 times a week over 2 weeks (a total of 60 minutes). Acceptability was assessed through the frequency and duration of use (analytics analyzed at the end of the 2-week intervention period and at weeks 4 and 6) and through participant reports. The Child and Adolescent Mindfulness Measure (CAMM), General Help-Seeking Questionnaire (GHSQ), Flourishing Scale (FS), and Revised Children’s Anxiety and Depression Scale (RCADS) were assessed at baseline and week 2 to indicate possible effects. Focus groups were held in week 4.

Results: Most participants (n=39, 87%) used Match Emoji for at least 60 minutes over the 2-week intervention, with 80% (36/45) continuing to play the game after the intervention period. Mean change (from baseline to 2 weeks) on each measure was 1.38 (95% CI –0.03 to 2.79; \( P=0.06 \)) for CAMM; 0.8 (95% CI –2.71 to 4.31; \( P=0.64 \)) for GHSQ; –1.09 (95% CI –2.83 to 0.66; \( P=0.21 \)) for FS; and –3.42 (95% CI –6.84 to –0.001; \( P=0.49 \)) for RCADS. Focus group feedback suggested that Match Emoji was enjoyable and helpful.

Conclusions: The casual video game with mental health content appeared to be acceptable and provided a promising indication of possible therapeutic effects. This approach is worthy of further investigation.

International Registered Report Identifier (IRRID): RR2-10.2196/31588

(JMIR Serious Games 2024;12:e46697) doi:10.2196/46697

KEYWORDS
adolescent; anxiety; casual video games; digital mental health interventions; gaming; mental health; micro interventions; serious game; teenage; video game; youth

Introduction

Mental distress and low well-being are common among adolescents [1-3] and appear to have increased over the past decade, at least in high-income nations [4-6]. Cognitive behavioral therapy (CBT) and psychotropic medications are recommended for young people experiencing mental health disorders [7,8]. Preventative programs that aim to buffer against higher levels of distress later in life also exist for young people.
Digital mental health interventions (DMHIs) refer to specialized content, support, or therapy for mental health conditions delivered electronically to treat, alleviate, or manage symptoms [11]. DMHIs encompass various technologies, including computerized CBT programs, chatbots, virtual reality for mental health conditions, games for mental health, apps, and interactive web pages [12]. Systematic reviews have shown promising effects for specific DMHIs across various age groups [13,14], such as CBT therapies for anxiety and depression [15,16]. Quality DMHIs can address some of the challenges often impacting face-to-face treatments [17,18]. For example, well-designed DMHIs can be used by young people irrespective of their level of distress, and they can be scaled up at a low cost due to their reduced reliance on clinically trained professionals [19,20].

Although this method of delivering mental health content is promising, engagement with DMHIs outside trial settings is typically lower than in trials [16]. Even playful interventions, such as Pesky gNATs [21] and SPARX [22], designed to appeal to young people’s interest in computer games, have had limited evidence of engagement [16]. In part, these findings may reflect mismatches between how end users engage with technology and the way tools are provided (e.g., sessions approximating weekly face-to-face therapies may be a poor match with contemporary patterns of internet use) [16,23,24]. Moreover, a lack of appealing options, lack of trust, or uncertainty about digital tools for mental health purposes may create additional barriers [22-24]. Therefore, while DMHIs have a great capacity to address mental health needs, it is important to keep exploring new opportunities to improve engagement [19,25,26].

Casual video games (CVGs) refer to simple games that can be played in short bursts of time, require no specialized skills, are often used for relaxation [27] and distraction purposes [28], and are generally free or low-cost to download and play. Well-known CVGs include “Rise Up” and “Angry Birds.” “Rise Up” has been downloaded over 10 million times on the Google Play Store worldwide, and “Angry Birds” is played for approximately 200 million minutes daily [29,30]. Given their popularity and potential therapeutic effects, CVGs may be an approach that could be explored for delivering mental health and well-being content [28].

We systematically reviewed the effects of CVGs on anxiety, depression, stress, and low mood [31]. We found that 12 of the 13 trials reported promising results on their respective outcome measures. Following this work, we developed simple prototypes of CVGs with mental health concepts based on the puzzle, word, and match-3 genres and reviewed these in focus groups and interviews with young adolescents [32]. Young people indicated interest in this idea, with a match-3-style CVG being preferred. Subsequently, a game designer was contracted to develop the first version of Match Emoji, a simple match-3 CVG that includes brief text-based mental health and well-being messages, which have been previously described [33]. In brief, this includes short “micromessages,” which were developed using psychological well-being literature and were sometimes linked to gameplay, for example, “Great job focusing and matching the emojis!” and “Phew! Take a short breath to help focus again.” Subsequently, we held think-aloud interviews [34] with a small group of young adolescents to refine components.

In this study, we aimed to conduct a small open trial to explore the feasibility of using Match Emoji to strengthen the mental health and well-being of adolescents in a school setting. Findings from the study can help develop the literature on this new possible method for delivering mental health and inform processes for a possible future randomized controlled trial.

Methods

Design

The recruitment procedures, sample size, and analyses differed from those planned and published in our protocol paper [35] due to COVID-19 pandemic–related restrictions. Each departure from protocol is documented in the relevant section below.

This feasibility study used a mixed methods design. Adolescents attending New Zealand intermediate and high schools were recruited to participate in this study. They were shown how to use Match Emoji and then asked to play for 15 minutes, 2-3 times a week over 2 weeks (a total of 60 minutes). Analysis of game use, analytic data, and focus groups were held with all participants to explore the acceptability of Match Emoji. The therapeutic potential of the game was assessed by changes in mental health and well-being, which were assessed by 4 validated instruments.

Recruitment

Before the onset of the COVID-19 pandemic, we developed a protocol to outline the guidelines for conducting the trial, including how participants would be recruited [35]. Initially, as 1 local secondary school had expressed interest in participating, we aimed to recruit students between the ages of 13 and 15 years from this school across 2-4 classrooms. However, several teachers had become ill during the recruitment phase, and the secondary school could no longer participate in the study. As such, we approached 2 secondary schools (students aged 12-18 years) and an intermediate school (students aged 12-14 years), which all expressed interest in participating in the study.

In the secondary school, we described the study to an assembly of over 400 students in years 9 and 10 (aged 12-14 years). Those interested in participating in the trial and with access to a smartphone or tablet were asked to take home information, an assent form, and a consent form for their parent or guardian. Of the 42 interested students, only 6 returned both forms. When recruiting participants in each intermediate school, the New Zealand government implemented restrictions on indoor face-to-face gatherings. At this time, indoor gatherings of up to 100 people were allowed. As such, instead of recruiting participants in an assembly, we delivered a 10- to 15-minute face-to-face presentation to students in each classroom, explaining the theory and research underpinning Match Emoji. In total, 39 returned the assent and parental consent forms. Given the primary aims, the inclusion criteria were students aged between 12 and 14 years who had access to a phone that could...
download Match Emoji and provided written consent from a parent or caregiver.

**Study Procedure**

Consenting participants completed the Child and Adolescent Mindfulness Measure (CAMM), General Help-Seeking Questionnaire (GHSQ), Flourishing Scale (FS), and Revised Children’s Anxiety and Depression Scale (RCADS) at baseline. These assessments were completed in groups of 6 in the high school and 30 in the intermediate school. Students completed the questionnaires at their desks and were separated at least a meter apart from each other to protect privacy. Instructions on how to play and install the game were provided, and participants were given an opportunity to ask questions directly or email the primary researcher. Next, they were asked to play Match Emoji 2-3 times a week for a minimum of 15 minutes per session for 2 weeks (ie, a minimum of 60 minutes in total). Questions were repeated after the 2-week intervention period. All participants were invited to participate in focus groups held at each school 2 weeks later (4 weeks after the study began). After the study, *koha* (food and drink) was provided to acknowledge the student’s effort. No financial incentives or gifts were provided.

**The Intervention**

The Match Emoji rationale, content, and processes have previously been described [33]. In brief, the micromessages in Match Emoji are based upon psychological well-being literature, specifically the *Five Ways to Wellbeing* [36]. As seen in Figure 1, these messages appear instead of in-game advertisements and function as prompts. For example, players are encouraged to read the message and practice skills including diaphragmatic breathing, noticing thoughts, or normalizing difficult emotions. In terms of the gameplay, users must identify and match 3 or more similar colored emojis together in rows or squares (a “match-3” game) to earn points. There are 6 different colored and shaped emojis, each representing an emotion or idea. The game has 99 levels, each designed to be completed within a few minutes, with a player advancing to the next level on completion of the current level. The gameplay becomes increasingly challenging as the player progresses.

**Figure 1.** Screenshot of the Match Emoji video game.

**Measures and Outcomes**

Demographic data were collected at baseline. Students who reported more than 1 ethnicity were categorized using the New Zealand Census ethnicity prioritization method [37].

Acceptability was assessed by the proportion of approached schools who agreed to participate, the number of participants who are able to download the game on their phone and those who fully participated in the study, and student feedback in focus groups. At week 4 (ie, after the intervention period), all participants were invited to take part in a 45-minute focus group at their school to explore their views of the intervention. Questions included (1) What parts of the game did you like? (2) What parts of the game could be improved? (3) What did you learn from playing the game? (4) Did you try and use any of the ideas from the game, and if so, which ones? and (5) Do you think you will continue to play Match Emoji? A general inductive approach was used to analyze the data from the focus groups [38]. The first author (RP) read participants’ responses several times to identify emerging themes and categories from the raw data. A research assistant read through the raw data to ensure the themes reflected the essence of the category. Appropriate quotes that conveyed the key core themes were recorded and integrated into the results. Lastly, game analytics for minutes played and the number of sessions were recorded on the Unity platform [39]. Unity is a secure platform for creating and operating interactive games.
The secondary outcome measures assessing therapeutic potential were changes from the pre- to postintervention (baseline and 2 weeks) time period on mental health and well-being domains calculated from the CAMM, a 10-item instrument measuring acceptance and mindfulness for use with children and adolescents aged 10 between 17 years; the GHSQ, which measures formal help-seeking intentions for nonsuicidal and suicidal problems; the 8-item FS, which measures self-perceived success in important areas such as relationships, self-esteem, purpose, and optimism as a single psychological well-being score; and the RCADS, a 47-item youth self-report questionnaire with subscales, such as separation anxiety disorder and generalized anxiety disorder.

The specific mental health and well-being domains assessed were mindfulness derived from the CAMM, help-seeking from the GHSQ, psychological well-being from the FS, and overall anxiety and depression score from the RCAD. Pretest and posttest summary statistics (mean, median, range, and SD) were computed using the R software (R Foundation for Statistical Computing) developer package. Data were assessed for normality using the Shapiro-Wilk normality test. Since data were not normally distributed, the nonparametric Wilcoxon signed rank test was used to compare the means between pairs of values (pre and post).

Ethical Considerations
This study received ethics approval from the New Zealand Health and Disability Ethics Committee (21/NTA/34) on May 28, 2021. Data was de-identified and all participants provided informed consent. No financial compensation was provided to the study participants.

Results
Participants
Of the 45 adolescents who participated in the study (mean age 12.5, SD 0.33; range 12-14 years), 26 (57%) were New Zealand European, 12 (26%) were Māori, and 7 (15%) were Asian or Pacific. As seen in Table 1, the majority (n=27, 60%) were boys, while 15 (33%) were girls and 3 (6%) were nonbinary.

Table 1. Demographics of participants (N=45).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>12.5 (0.33)</td>
</tr>
<tr>
<td>Range</td>
<td>12-14</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Boy</td>
<td>27 (60)</td>
</tr>
<tr>
<td>Girl</td>
<td>15 (33)</td>
</tr>
<tr>
<td>Nonbinary</td>
<td>3 (6)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Māori</td>
<td>12 (26)</td>
</tr>
<tr>
<td>New Zealand European</td>
<td>26 (57)</td>
</tr>
<tr>
<td>Pacific</td>
<td>4 (9)</td>
</tr>
</tbody>
</table>

Acceptability
On average, each participant played 7.5 sessions for 24 minutes across the 2 weeks, comprising 180 minutes in total. In addition, data recorded from the focus groups suggested that, on average, participants completed 50 out of the 99 available levels during the 2-week duration of the study. Most participants in the focus groups said they would continue playing the game after completing the study. A total of 38 (84%) participants said they “would” continue to play the game, while 5 (11%) said they “might” continue to play. Only 2 (4%) participants said they would not continue to play the game after the trial. In addition, 36 (80%) reported playing Match Emoji after week 4, and 32 (71%) were still playing after week 6, according to game analytics from the Unity Platform. Findings from the focus groups suggested that participants enjoyed playing Match Emoji for several reasons. First, participants enjoyed the convenience of the game. For instance, many participants reported playing Match Emoji across multiple environments, including waiting rooms at the dentist, bus stops, and long car rides. As no internet connection was needed, participants could access the game whenever they wished. A participant explained, “I could play the game even when there was no Wi-Fi,” while another said, “The game was really good when waiting for appointments (be)cause it could distract me for a bit and didn’t use up data.” Second, many participants reported enjoyment from playing the game. They described this enjoyment as stemming from game features such as increasing levels of challenge, the variety of emojis, and clear goals: “it was fun (be)cause the game got harder, but you knew what you had to do.” While there was some level of challenge, the simplicity of the game allowed students to bypass traditional barriers to CVGs, such as instructional videos. One participant described Match Emoji as a “super easy game to understand and play.” A smaller group of participants also provided suggestions about game features. This group appeared to be more frequent users of CVGs, as they provided recommendations based on other games they had

https://games.jmir.org/2024/1/e46697

JMJ Serious Games 2024 | vol. 12 | e46697 | p.38
(page number not for citation purposes)
played. One participant suggested, “You could add more rewards or more characters and then get more power-ups like Fortnite,” while another recommended, “coins, customization, themed music, and bonus rounds... add stuff like they have in other casual games.”

In general, participants liked the subtle aspect of accessing mental health content. As 1 participant mentioned, “the messages are a nice way of getting mental health information out there that isn’t in your face.” There was a high consensus that they preferred micromessages over typical in-game advertisements. However, some were initially skeptical about their value, “the messages were cringe at first but got way better.”

Of the intermediate and secondary schools approached to participate in the research, only 3 (25%) of the 12 took part in the study. Only 3 (7%) out of the 45 participants could not download Match Emoji onto their phones. In each case, this was because their phone had limited capacity to download the necessary software. All participants completed baseline and follow-up assessments, but several needed clarifications on wording related to the RCADS questionnaire items.

### Indicators of Possible Effects

As seen in Table 2, a small positive change was observed on the CAMM (mean difference 1.38, 95% CI –0.03 to 2.79) and on the RCADS (mean difference 3.42, 95% CI –6.84 to –0.001). In focus groups, when asked, “What did you learn from playing the game,” a number of participants answered that playing the game was helpful for their mental health and well-being: “I reckon playing the game for a bit of time was helpful for my mental health (be)cause it took my mind of stuff.” When asked, “Did you try and use any of the ideas from the game, and if so, which ones?” Several participants reported using specific skills suggested in Match Emoji: “Once when I started to think about annoying stuff, I tried the breathing thing, and it was actually pretty helpful,” and “I remember I got pretty mad at my brother and used the noticing a thought approach.”

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Baseline, mean (SD)</th>
<th>Postintervention, mean (SD)</th>
<th>Mean differences (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMM*</td>
<td>22.44 (8.35)</td>
<td>23.82 (8.93)</td>
<td>1.38 (–0.03 to 2.79)</td>
<td>.06</td>
</tr>
<tr>
<td>GHSQ**</td>
<td>62.89 (21.96)</td>
<td>63.69 (23.30)</td>
<td>0.8 (–2.71 to 4.31)</td>
<td>.65</td>
</tr>
<tr>
<td>FS†</td>
<td>41.71 (11.58)</td>
<td>40.62 (12.07)</td>
<td>–1.09 (–2.83; 0.66)</td>
<td>.22</td>
</tr>
<tr>
<td>RCADS‡</td>
<td>46.24 (26.39)</td>
<td>42.82 (26.49)</td>
<td>–3.42 (–6.84 to –0.001)</td>
<td>.049</td>
</tr>
</tbody>
</table>

*aCAMM: Child and Adolescent Mindfulness Measure (mindfulness).
**GHSQ: General Help-Seeking Questionnaire (help-seeking).
†FS: Flourishing Scale (psychological well-being).
‡RCADS: Revised Children’s anxiety and Depression Scale (overall anxiety and depression).

### Discussion

#### Overview

In this study, we found that a CVG with psychological well-being concepts (Match Emoji) was a new and engaging mechanism of change that provided a promising indication of possible therapeutic impact. Most participants played more often and for a longer period than was requested for the study. Indeed, most participants continued to play in week 4. Small improvements in mindfulness assessed by CAMM and a small decrease in overall anxiety and depression assessed by RCADS were recorded. Given these promising changes, participants may have learned skills related to reducing their level of anxiety through playing Match Emoji. The findings of this small open feasibility trial indicate that the Match Emoji CVG was an acceptable way to support mental health and well-being in adolescents aged between 12 and 14 years.

Participants reported a high level of acceptability with Match Emoji, as evidenced by the game analytics, qualitative feedback, and the large portion of participants who were still playing the game in weeks 4 and 6. The percentage of participants who stated they continued to play Match Emoji even after week 6 of the study (n=32, 71%) is contrary to the poor retention rate typically found across the range of digital interventions.

Real-world data on user engagement with popular mental health apps suggest that a small portion of users stay engaged with digital health interventions [16]. For example, once a health app is downloaded, approximately 4% of users continue to use the app after 15 days [24]. It is possible that the ongoing consultation with end users from the beginning of the development of Match Emoji, the simplicity with which CVGs can be played “on the go,” across environments with no Wi-Fi, and how playing CVGs fits with adolescents’ current behavior patterns may have been attributed to the high level of acceptability and engagement. That is, as many adolescents already play CVGs [32], there is less effort required to learn and change existing ways of engaging with technology. Data from the focus groups corroborated these findings. More specifically, participants mentioned they enjoyed playing for short periods across environments in comparison to computer games or those mobile phone games that require data to access. This is consistent with our previous work [32] and research [19], which suggests young adolescents tend to prefer brief therapeutic encounters. Moreover, Match Emoji enables large portions of the population to receive the same content irrespective of their proficiency with gaming or access to the internet, addressing a significant barrier to equity and engagement with DMHIs [11].
Our finding that most participants preferred micromessages over typical in-game advertisements is consistent with research assessing how in-game advertising in the form of short videos is distracting and can lead to disengagement, particularly among young people who often have a relatively short attention span \[40,41\]. Although paid versions can avoid advertisements, young people are reluctant to pay for them \[41\]. Thus, Match Emoji represents an opportunity for public health interventions to provide appealing free CVGs that replace the advertising with health-related micromessaging that is not distracting, annoying, or potentially harmful, as is the case with in-game advertising.

Similarly, diverse preferences were found with gamification elements of Match Emoji. The various preferences toward micromessages and game features among participants are consistent with the literature that suggests adolescents have different opinions about the type of DMHI they are attracted to \[42,43\]. Thus, while some adolescents may be frequent CVG users and interested in gamification elements, others may be less focused on these features and more attracted to learning about mental health and well-being \[44\]. In essence, opportunities to embed therapeutic processes within game elements are plentiful when researchers and game developers collaborate and are creative.

The protocol and implementation of this study were completed during the COVID-19–lined social distancing practices, which resulted in frequent changes to the restrictions on the size of inside gatherings and how educational facilities operated. Apart from the implications of the pandemic, 3 participants in the study could not download Match Emoji. This was because their phones lacked the storage capacity required to download the latest software and the game. Future research could use methods to compress digital mental health apps such as Match Emoji. In this way, the size of the app may better align with the capacity of users’ technology. In addition, some participants struggled to understand several questions on the RCADS. These questions were discussed in more detail with each participant to ensure they understood the meaning of each one. Despite these challenges, no significant issues occurred with conducting the study in a primary and intermediate school context.

**Limitations**

There are limitations to this study; these include departures from the protocol due to COVID-19 impacts, which resulted in a small exploratory open trial only. There were also limited resources to conduct the study; this meant that the first author (RP) introduced the game to participants, led the recruitment process, supported the completion of assessments before and after playing the game, and facilitated the discussions about the game. Thus, the interpretation of the students’ feedback could be overly positive, and participants’ opinions and thoughts could be influenced by social pressures, including normative and informative conformity.

Further, the self-assessment outcome measures relied on the comprehension skills of young participants. While some participants raised their hands when unsure of a question, others may have merely guessed. Nevertheless, the 4 assessments appeared to be easy to implement in a short amount of time. Third, when recruiting participants at the secondary school, only 6 (14%) out of the 42 participants who signed the assent form returned their parental consent form for reasons unknown, suggesting that a different process is needed to recruit older adolescents for future trials. Lastly, students were not recruited based on their level of mental distress. Therefore, the results may have been affected by floor effects, whereby their mental health and well-being scores were already good or optimal and thus unlikely to improve any further. Despite these challenges and preliminary results, these findings are of interest as this is the first study to assess the feasibility of a co-designed CVG with psychological well-being concepts.

**Conclusion**

Findings from this feasibility study suggest that Match Emoji, the purpose-built CVG with brief mental health messages, is promising as an acceptable and feasible approach for young adolescents. Future research should test clinical impacts through a randomized controlled trial. More broadly, the research also highlights the possibility of CVGs as a novel mechanism of delivering simple mental health and well-being messages.

**Acknowledgments**

We would like to acknowledge the young people who participated in the study and the teachers who helped with the recruitment process. This study was funded by Cure Kids (innovation seed fund: 3918).

**Data Availability**

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

**Authors’ Contributions**

RP and TF were involved in the conceptualization of the game and study. LTM and JM were involved in the methodology and formal analysis. All authors were involved in writing the original draft and reviewing and editing the paper. All authors have read and agreed to the published version of the manuscript.
Conflicts of Interest

TF is a co-developer of SPARX, a computerized cognitive behavioral therapy program for adolescent depression. The intellectual property for SPARX is owned by Uniservices at the University of Auckland, and co-developers can benefit financially from the licensing of SPARX outside of New Zealand.

References


Abbreviations

- CAMM: Child and Adolescent Mindfulness Measure
- CBT: cognitive behavioral therapy
- CVG: casual video game
- DMHI: digital mental health intervention
- FS: Flourishing Scale
- GHSQ: General Health-Seeking Questionnaire
- RCADS: Revised Children’s Anxiety and Depression Scale
A Smartphone-Gamified Virtual Reality Exposure Therapy Augmented With Biofeedback for Ailurophobia: Development and Evaluation Study

Ali Khaleghi¹, PhD; Abbas Narimani², MSc; Zahra Aghaei³, PhD; Anahita Khorrami Banaraki⁴,⁵, MD, PhD; Peyman Hassani-Abharian⁴,⁵, MD, PhD

¹Iranian Light Source Facility, Institute for Research in Fundamental Sciences, Tehran, Iran
²Department of Computer Engineering, Imam Khomeini International University, Qazvin, Iran
³Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran
⁴Brain and Cognition Clinic, Institute for Cognitive Science Studies, Tehran, Iran
⁵Department of Cognitive Psychology and Rehabilitation, Institute for Cognitive Science Studies, Tehran, Iran

*these authors contributed equally

Corresponding Author:
Ali Khaleghi, PhD
Iranian Light Source Facility
Institute for Research in Fundamental Sciences
Opposite the Araj, Artesh Highway, Aghdassieh
Tehran, 19395-5746
Iran
Phone: 98 9121003006
Email: ali.khaleghi.ir@gmail.com

Abstract

Background: To the best of our knowledge, no specialized research has been conducted to address ailurophobia (fear of cats) in Iran or globally. This has driven our project, along with the prevalence of ailurophobia and the absence of a gamified virtual reality exposure therapy (VRET) that incorporates affordable and easily accessible biofeedback (BF) tools. We hypothesize that a gamified VRET augmented with BF will yield more positive effects than a similar device lacking BF.

Objective: This study primarily focuses on the development and preliminary evaluation of a smartphone-gamified VRET integrated with BF, targeting animal phobia, with a specific case study on ailurophobia. The secondary objectives are using affordable and readily available BF found in devices such as smart bands and smartwatches and creating a mobile virtual reality gamified app to improve patients’ adherence to treatments while simultaneously enhancing the app’s accessibility, scalability, and outreach.

Methods: Evaluations encompassed 3 methods. First, we identified the tool’s potential positive effects on phobia interventions, exploring 4 effects: intrinsic motivation, simulation of fearful situations, management of stressful circumstances without therapists’ presence and mitigation of catastrophic thoughts, and preliminary effects on ailurophobia treatment. The secondary objectives are using affordable and readily available BF found in devices such as smart bands and smartwatches and creating a mobile virtual reality gamified app to improve patients’ adherence to treatments while simultaneously enhancing the app’s accessibility, scalability, and outreach.

Results: The smartphone-gamified VRET augmented with BF exhibited better results on the identified effects compared with the non-BF version and contributed to normalizing encounters with cats. Moreover, 41 of the 44 heuristics achieved a percentage above 62%, indicating its potential as a therapeutic product and its ability to enhance patient adherence to treatments. Patient preferences on the treatment and its strengths and weaknesses were provided for further improvement.

Conclusions: The tool has the potential to evolve into a comprehensive solution by incorporating various types of cats and their behaviors, simulating environments in which they are commonly found, and enhancing its appeal through an increased sense of adventure without inducing unrealistic fears. By adapting fear elements, the game can be tailored to treat various animal phobias.
Phobia-focused games should avoid action and combat scenarios to prevent reinforcement of fear responses. After rigorous evaluation, further exploration is required to provide remote use beyond clinical settings.

**KEYWORDS**

animal phobia; specific phobia; ailurophobia; cat phobia; biofeedback; smartphones; virtual reality; gamification; mobile phone

### Introduction

**Specific Phobia and Available Therapies**

Specific phobia is the most common anxiety disorder, with a lifetime prevalence of 12.5% [1]. It is characterized by an extreme and persistent fear of a specific object or situation [2], leading to substantial disruptions in daily life and heightened anxiety. Many individuals restructure their lives to evade their fears over extended periods [3,4]. Prolonged phobia detrimentally affects academic, social, and family aspects, compromising overall quality of life [5]. Situational (eg, fear of enclosed spaces and flying), natural environment (eg, fear of heights and storms), animal (eg, fears of snakes, spiders, and cats), and blood or injection or injury (eg, fears of medical procedures and seeing blood) fears are subtypes of specific phobias, with animal and natural environment phobias being more prevalent [3].

Phobia interventions are categorized into exposure therapies (eg, direct in vivo exposure, systematic desensitization, imaginal exposure, and virtual reality [VR]) and nonexposure approaches (eg, cognitive therapy and progressive muscle relaxation). There has been a trend toward adopting brief, intensive, or concentrated treatments to manage anxiety [5]. Among the available treatments, exposure therapies are the most commonly used approach for specific phobias [6]. However, although specific phobias are highly treatable, only 31% of patients seek treatment and, among those, only 43.4% seek mental health services [3]. Moreover, some patients might be unable to complete the treatment because of severe reactions, resulting in an attrition rate of 45% [3,7]. In total, 3 main factors contribute to this percentage [7]: (1) perceiving treatments as highly aversive and frightening; (2) the need to visit clinics throughout the treatment, causing relationship and ethical issues, such as perceived cruelty when therapists intentionally evoke fear; and (3) the lack of appealing treatments.

**Gamified VR Exposure Therapies Augmented With Biofeedback**

To overcome the limitations of exposure therapy methods, incorporating new technologies becomes imperative. Gamification, VR, and biofeedback (BF) are promising options. However, our research indicates that few studies have simultaneously used these technologies for specific phobias. Virtual reality exposure therapy (VRET) uses 360° computer-generated simulations [8,9] similar to traditional exposure therapies [2]. Meta-analyses have shown that VRETs are effective and their performance can rival standard exposure therapies [2]. VR’s application in cognitive impairment, anxiety disorders, pain management, phobias, posttraumatic stress disorder, rehabilitation, and eating disorders, among others, has surged because of its immersive realism [8,10]. To treat phobias, VR is a safer, less embarrassing, and cost-effective solution by simulating fear-inducing situations in a controlled environment [8,9,11]. However, VR alone may not address all exposure therapy disadvantages, and enhancing the attractiveness of VRETs is crucial for treatment success. Researchers have explored the potential of gamified VRETs in treating phobias [2,12,13]. Gamification, a strategy derived from video game–based approaches, has proven successful in achieving serious objectives across various fields, including the workplace [14], education [15], marketing [16], mental health [17-19], learning disabilities [20,21], and lazy eye treatment [22]. The primary inherent feature of digital games is their high-level motivational potential [23]. Video games’ appeal, engagement, and effectiveness encourage players and frequent use [18]. Attractiveness is beneficial for overcoming people’s reluctance to seek treatment, broadening the reach of gamified interventions [18]. The engaging nature of gamification enhances users’ experiences, as players are driven to win, explore stories, and ultimately reduce attrition rates [12,18,24]. The effectiveness aspect offers opportunities for achieving serious objectives such as behavior changes [18]. In a gamified product, elements such as scores, badges, and levels are integrated from games into nongame contexts, while not necessarily offering a complete gaming experience [18,25].

Human emotion recognition sensors or BF is another technology that can enhance gamified interventions. This technology serves 2 crucial purposes. First, it boosts their level of attractiveness by leveraging a strategy commonly used in video games to increase engagement [26]. Second, it addresses some of the limitations of traditional methods by potentially reducing or eliminating the need for therapists’ constant presence. These sensors work by measuring various body parameters or electrical impulses in the nervous system to identify different emotions and track their changes [27]. Common techniques include electroencephalography, skin resistance measurements, blood pressure, heart rate (HR), eye activity, and motion analysis. With advancements in chipset manufacturing, BF has become more accessible, portable, efficient, and affordable. Users can easily access their data, thereby enabling self-regulation and monitoring. These technologies are incorporated into smart wristbands and watches to help individuals regulate anxiety in their daily lives. BF therapies have shown positive effects in treating conditions such as migraines [28] and attention-deficit/hyperactivity disorder in children [29].

**Objectives**

The primary aim of this study was to develop and conduct a preliminary assessment of smartphone-gamified VRET augmented with BF for the treatment of cat phobia (ailurophobia). We hypothesize that this tool will outperform
gamified VRET without BF in various aspects. Limited evidence exists on animal phobia in Iran, particularly ailurophobia. Observations at the Cognitive and Brain Clinic in Tehran revealed a substantial prevalence of this phobia, as reported by the fourth author, who is a cognitive expert and psychologist attending to cat phobia patients daily. Owing to the abundance of cats in most Iranian cities, encounters are inevitable, resulting in daily challenges for patients walking on the streets and alleys. The secondary objectives were as follows:

1. Using affordable and accessible BF tools in devices such as smart bands and smartwatches to serve as both BF and a game mechanic, enhancing engagement and efficacy.
2. Developing a mobile VR–gamified app to enhance patients’ adherence to phobia treatment and expand the app’s accessibility, scale, and reach.

To evaluate the effectiveness of the tool, its potential positive effects on phobia interventions were examined. The tool’s impact on the effects was examined by dividing the participants into BF and non-BF groups. In addition, we considered the playability and usability aspects of the tool, along with patients’ preferences, to optimize its performance and enhance usability for future improvements.

Methods

Design and Development

Our primary objective was to present fear elements indirectly to the player, ensuring that interacting or not interacting with them would not affect the game’s progress. The secondary objective was to create a general game design model that could be easily customized for specific phobias, particularly animal phobias. During the initial game development meeting, 2 game design experts (a game designer and a gamification expert) collaborated with 2 cognitive science experts (one of whom also specialized in cognitive games). They engaged in a discussion regarding the essential components required to simulate stress. Size, color, and behavior of the stimuli were introduced as fundamental elements for replicating the desired scenarios. The game team then devised the game stages using a maze design. In the second meeting, cognitive experts suggested simplifying the design to accommodate players of all ages. As it involved memory and problem-solving, it was rejected, leading to a more straightforward game plan that focused on finding lost objects in a park. In the third session, this plan received approval and was tested on a woman aged 40 years with cat phobia, who was selected based on her self-reported fear of cats. She stated, “I experienced a lot of fear during playing.” In the fourth session, minor adjustments were made to the game. In total, 2 psychologists from Tehran University found the initial voice of the guide annoying, thereby hindering patient motivation. The overall view of the sessions is presented in Figure 1.

Our game’s storyline was inspired by the “Hot and Cold” game. One group hides an object, and the other group should find it using verbal clues such as “colder” as they move away and “hotter” as they get closer. The experience is similar to that of a park with diverse paths. Players are on a quest to discover diamonds concealed within treasure boxes, all while walking along these pathways. Each game session comprises 4 short yet consecutive levels. At each level, players must determine their distance from each box by perceiving changes in the sound consistently played. Moreover, a hint ribbon shows the player’s distance to the box for increased engagement. After locating the box, players must stay in front of it for a specific duration to open it, with the time increasing at later levels. The players must open the previous level’s box to unlock the next challenge.

Regarding authors’ concerns about spreading the game to individuals with phobias, smartphones were chosen as the primary platform. Using smartphones as a VR tool requires affordable mobile VR glasses, which are significantly cheaper than other options such as Oculus or HTC VR. The primary challenge in mobile VR is the user interaction limitations. The game uses Gaze, a pointer on the screen that allows users to interact through head movements, thus providing a mouselike experience. In addition, the game incorporates joysticks connected to the phone, thus offering more interactive possibilities.

The intensity of the fear elements must be balanced based on the game’s progress and levels, as in previous studies [7,13]. The escalation of fear stimuli is determined by the following features, each with its own difficulty level. Moreover, these elements can be further amplified in tandem with player’s advancement.

1. Visual elements: the fear-triggering elements include cat photos, fantasy cat models, low-poly cat models with minimal details, and high-poly cat models that closely resemble real cats. According to experts, individuals who fear something may also react to objects and shapes that resemble it. For instance, people who are afraid of cats might experience fear when encountering a cat picture or a furry object. This phenomenon is directly related to the degree and intensity of the individual’s fear [30]. Figure 2 illustrates the game environment.

2. Fear elements’ sound: the scary elements vary from silent to those with terrifying sounds. In intense situations, cats produce specific sounds that could heighten anxiety. The timing of when the sound is played also adds to the diversity. For example, when players are near a cat, the sounds it emits could intensify their fear.

4. The quality of fear elements’ behaviors: studying the behavior quality of a stimulus is under investigation [31,32]. A cat jumping from one point to another evokes more fear than a cat simply standing still. Various animations were designed for 3D model cats. The fantastical cat playfully turns its head and randomly spins around. The low-poly cat remains stationary, solely turning its head. In the final level, the high-poly cat features 3 different animations. The first 2 animations portrayed the cat at rest, either shaking its tail and head or cleaning its paws. The third animation involves the cats’ walking behavior.

5. Interactable elements: fear elements that respond to the player’s presence add a sense of authenticity to the game, elevating immersion and allowing for anxiety manipulation. Cats may react by turning, approaching, or fleeing when a player gets closer. Both low-poly and high-poly cats respond to the
player’s presence. The manner in which the elements react was also classified. Although the fantasy cat remains unresponsive, the low-poly cat acknowledges players by turning their heads and looking at them when they enter the zone. At the last level, the realistic cat not only faces the players but also follows them until they exit from its zone.

6. Fear elements’ size: element size could also amplify fear. In the final level, some cats are larger, preparing players to confront more intimidating situations.

7. Fear elements’ numbers: seeing numerous cats creates a feeling of being surrounded, indirectly encouraging players to confront their fear. As players approach the boxes, fear intensifies, peaking around those areas.

The quantity and type of elements can be customized based on players’ preferences and conditions (Figure 3), which is beneficial when they need to concentrate on a specific scary element. Furthermore, a player who does not fear an element can eliminate it from the game.

A VR Android game was developed using the Unity game engine, incorporating the Amazfit Bip smartwatch. In anxiety treatments, HR variability is a common BF technique for stress management [33]. However, because of limitations in receiving these signals through conventional smartwatches and wrist bands, HR was used instead of HR variability. HR data are accessible in smartwatches through Bluetooth low energy technology [33]. A plug-in for the Unity3D game engine was implemented to integrate smartwatch data into the game. The player’s HR was incorporated into the experience as a game mechanic. The HR was displayed on the corner of the screen. A total of three BF techniques were used in this study: (1) displaying changes in players’ bodies to inform and manage anxiety [33]; (2) keeping HR within specific limits allows players to earn the game’s prize, a diamond, promoting relaxation skills for stressful situations [34]; and (3) maintaining a low HR for a period allowed players to open boxes and collect more diamonds [33,35].

Figure 1. The overall view of sessions between the game designers and other related experts.
Trial Design, Participants, and Procedure

**Overview**

In total, three methods were used to evaluate the tool: (1) identifying its potential positive effects it could have in phobia interventions. To assess the game’s impact on these effects, participants were divided into BF and non-BF groups, with the only differences being the use of smartwatches; (2) gathering user preferences about the treatment; and (3) considering the tool’s playability and usability aspects for subsequent optimization and improved usability.

**Ethical Considerations**

This study was approved by the Research Ethics Committees of the Institute for Cognitive Science Studies (IR.UT. IRICSS.REC.1401.047). Informed consent was obtained from participants. They had the freedom to withdraw from the study at any time. The participants’ data were anonymized. To compensate for time, participants were informed that a smartphone-compatible version of the game would be provided free after its finalization.

**Participants**

The snowball method was used for recruitments. One attractive advertisement was prepared in Farsi and shared within various working, educational, and family groups on Instagram and
WhatsApp. Receivers were asked to help by sharing the advertisement with their own groups. Recruitment took place from September 8 to October 14, 2022, in 2 provinces in Iran: Lorestan and Tehran. Each test session lasted up to 3 hours, and the participants had the flexibility to choose the test location. Random assignment was used to allocate the participants to the study arms.

Inclusion criteria were (1) providing informed written consent, (2) understanding and reading Persian, and (3) scoring ≥55 on the Fear of Cats Questionnaire (FCQ). Exclusion criteria were (1) currently receiving psychological treatment for ailurophobia; (2) having another severe mental disorder (alcohol or substance abuse, psychotic disorder, dementia, or bipolar disorder); (3) diagnosed with a severe personality disorder; (4) experiencing depressive symptoms or suicidal ideation; (5) heart disease; (6) vision or balance problems affecting the VR experience; (7) pregnancies exceeding 3 months; and (7) fear of cats only in a few and exceptional cases. An image of participants is presented in Figure 4.

Figure 4. Depiction of participants.

Identifying the Positive Effects of the Gamified VRET Augmented With BF

Effect 1: Intrinsic Motivation

One primary positive effect that the app could have on phobia interventions is its ability to enhance intrinsic motivation. By incorporating gamification, VR, and BF, the app effectively motivated patients to actively engage in their treatment. We hypothesize that combining gamified VRET with BF will significantly increase motivation compared with a similar tool without BF.

To assess their impact on intrinsic motivation, we used subjective and objective measures. After each level, participants completed a 10-item questionnaire that was previously used to evaluate subjective engagement [36-39].

Participants played a minimum of 5 levels and completed the intrinsic motivation questionnaire after each level, except for the first. First-level data were excluded because of participants’ unfamiliarity with the experience. For the training step, the picture level of the game with 8 cat pictures was predetermined. The other settings regarding the type and quantity of stimuli for mandatory games are as follows:

- Game 1: fantasy model with 13 cats
- Game 2: low-poly model with 19 cats
- Games 3 and 4: high-poly model with 23 cats

We deliberately chose the last 2 steps in the same manner to examine the impact of repetitive tasks on the participants.

After the mandatory games, participants had the option to play the game for up to 4 additional times. During the voluntary sessions, participants were allowed to choose the type and number of cats, but the number of cats had to be selected in ascending order. In these sessions, we used a shorter version of the intrinsic motivation questionnaire with only 5 items, as used in Lumsden et al [36].

Effect 2: Simulating Fearful Situations

For phobia treatments to be effective, the game should evoke fear among individuals. To evaluate this, both groups were asked to rate their anxiety levels on a scale from 1 (“no anxiety”) to 10 (“extreme anxiety”) after any mandatory and voluntary sessions (except the first level).

Effect 3: Controlling Stressful Circumstances, Eliminating Therapists’ Presence, and Mitigating Catastrophic Thoughts

The game enables participants to implicitly learn relaxation techniques while confronting their fears. The box-opening mechanism involves standing in front of the box for gradually increasing durations. This combined approach, along with BF, has the potential to reduce the need for therapists’ presence.
After the experiments, the participants were asked two questions: (1) How well do you think you could manage your stress when dealing with a real cat after using the gamified app? (2) To what extent can our game eliminate the need for operators? The app’s attractive and fantasy environment was expected to alleviate catastrophic thoughts. Participants were also encouraged to share any positive signs of reducing their frequency of thinking about their fears.

**Effect 4: Preliminary Effects on Ailurophobia Treatment**

The study used before and after assessments with the State-Trait Anxiety Inventory (STAI) and FCQ to measure the game’s impact on phobia symptom changes. The STAI questionnaire comprises 40 questions, measuring state (S-scale) and trait (T-scale) anxiety using a 4-point Likert scale. Only the S-scale was used in this study. The evaluation of state anxiety can be used for any situation with a time interval determined by a researcher or a clinical specialist. Mahram developed the Persian version of the STAI, and its internal consistency was confirmed for the S- and T-scales (Cronbach α of .91 and .90, respectively) [40]. Another Iranian study also reported high reliability for the S- and T-scales with Cronbach α values of .93 and .90, respectively [41]. The FCQ questionnaire was derived from the Fear of Spiders Questionnaire (FSQ) to assess cat phobia, with all instances of the word “spider” replaced by “cat.” Furthermore, the question format was adjusted to suit the assessment of the cat phobia. The FSQ is an 18-item tool scored on a 7-point Likert-type scale to measure the level of spider phobia, yielding a total score ranging from 18 to 126. The FSQ demonstrates excellent internal consistency with Cronbach α ranging from .88 to .97 [42,43] and good test-retest reliability [42]. The FSQ has been used in previous studies for various phobias such as cockroaches [7,44,45], rats [46], and snakes [47].

After the games, each group of participants was instructed to play 2 levels of the game as the opposite group did. They were then asked to answer the following questions: (1) Which experiences do you prefer? (2) Which experiences had more novelty and were more attractive to you? (3) Which experience was more effective for improving your problem?

**Patients’ Preferences About the Designed Treatment**

To gather patients’ opinions on the implemented treatment, an adapted preference questionnaire [48] was used. This 6-item questionnaire focused on patients’ preferences regarding the types of cat models, their behaviors, sounds, and sizes. For example, the questions related to cat models are as follows: (1) If you could choose among the cat models, which one would you prefer? (2) Which cat model do you think would be more effective in helping you overcome your problem? (3) Which cat model do you find more logical for aiding in your progress? (4) Which of these cat models do you perceive as more aversive? (5) Which cat model would you recommend to a friend facing the same problem? (6) Are there any cat models missing in the game?

**Heuristics Evaluations**

The playability and usability aspects of the tool were examined through heuristic evaluations designed as semistructured interviews to optimize its performance and enhance usability. Participants completed a 5-Likert questionnaire covering user interfaces, VR experiences, BF, and game playability. Participants had the opportunity to provide additional comments. The evaluations incorporated 44 heuristics from studies [49-52]. We used Nielsen heuristics [49,50] to assess the interfaces, along with modified Nielsen principles for VR platforms [51].

In terms of game playability, a comprehensive evaluation was necessary to assess additional features, including gameplay, story, and mechanics, which went beyond simple interface usability evaluation [52]. We used the heuristic principles of playability introduced in [52], which carefully examine the various components of a game in terms of playability and enjoyment for the player, encompassing gameplay, mechanics, usability, and game story. In this study, we used the first 3 heuristics from this set.

**Statistical Analysis**

We evaluated the differences in subjective ratings of intrinsic motivation and levels of anxiety using ANOVA: 2-factor with replication of the total score, with session number as the time factor and task variant (the tool with and without BF) as the between-subjects factor. In addition, we used 1-way ANOVA with task variant as the between-subjects factor to investigate the effects of the tool on mitigating phobia symptoms. For analyzing the semistructured interviews, mean and SD scores were used.

**Results**

**Participants**

Of the 17 participants, 7 were excluded for (1) heart disease (n=1); (2) vision or balance problems (n=1; participants with VR-induced dizziness and severe nausea); (3) pregnancy (n=1); (4) personality disorders (n=1); and (5) fear of cats in specific situations (n=3; one was afraid of direct eye contact with cats, whereas 2 others were scared of black cats). Among the 10 included participants (Table 1), 1 individual had 2 other phobias: fear of public toilets (paruresis) and birds (ornithophobia), especially their beaks and legs. Another participant displayed general phobia of animals; even touching chicks elicited an electric shock response. In addition, the sight of cats, dogs (cynophobia) especially when they bark, and foxes caused annoyance and discomfort for her. Interestingly, she was more afraid of kittens than fully grown cats. Another patient had cynophobia and ailurophobia. Finally, 1 participant had a phobia of space and galaxies (to the extent of avoiding space-themed movies) as well as chicks phobia and ornithophobia stating, “I am even afraid of a bird in a cage that might come out and harm me.” This participant also avoided going to the park because of the fear of the animals. Given the prevalence of individuals experiencing multiple phobias, particularly fears related to various animals (zoophobia), such as cats, spiders, snakes, and dogs, it is crucial to explore the possibility of modifying the game to effectively address multiple types of phobias. The park environment appears to be conducive to addressing various animal phobias and specific phobias such as paruresis. Accessing 10 participants was hindered by the temporary filtering of Instagram in our country. In addition, 2 individuals declined to...
participate, expressing shyness and concerns about others noticing their phobia. Our observations suggest that men with ailurophobia conceal their fear more frequently. Notably, ailurophobia predominantly affected women, as 90% (9/10) of our participants were women (Table 1). Ailurophobia began in 70% (7/10) of the participants during childhood and 30% (3/10) during adolescence. The minimum and maximum ages of onset of phobia in the samples were 5 and 18 years, respectively. Regrettably, animal phobias in our country, particularly cat phobia, have been largely overlooked, leading individuals to live for many years in a completely curable condition without seeking treatment. Innovative and early interventions, for example, our tool, could treat patients from childhood when anxiety starts and reduce the negative impact of untreated phobias. There is a pressing need for screening and diagnostic games as a primary step, followed by therapeutic games. The main cause of participants’ phobia stemmed from an unexpected childhood encounter with a cat.

Table 1. Participants’ characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Non-BF^a</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>24 (7.31)</td>
<td>33.5 (7.16)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>5 (100)</td>
<td>4 (80)</td>
</tr>
<tr>
<td>Video game playing hours per week, mean (SD)</td>
<td>6 (8.52)</td>
<td>1.5 (3.08)</td>
</tr>
<tr>
<td>Median level of education</td>
<td>Diploma degree</td>
<td>Master’s degree</td>
</tr>
<tr>
<td>Years living with cat phobia, mean (SD)</td>
<td>15.6 (5.68)</td>
<td>21.6 (12.01)</td>
</tr>
<tr>
<td>Married, n (%)</td>
<td>2 (40)</td>
<td>2 (40)</td>
</tr>
<tr>
<td>The onset of phobia, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Childhood</td>
<td>4 (80)</td>
<td>3 (60)</td>
</tr>
<tr>
<td>Adolescence</td>
<td>1 (20)</td>
<td>2 (40)</td>
</tr>
<tr>
<td>Youth</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Adulthood</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Age of onset of phobia (years), range</td>
<td>9-13</td>
<td>5-18</td>
</tr>
</tbody>
</table>

^aBF: biofeedback.

Possible Positive Effects of the Gamified VRET Augmented With BF

Effect 1: Intrinsic Motivation

The average intrinsic motivation of the groups indicated better results for the BF group across all 4 mandatory games with 49 scores (the sum of motivation scores for BF vs non-BF in the first to fourth sessions were: 182 vs 174, 178 vs 169, 191 vs 160, and 182 vs 181). However, the results (P value [groups]=.15>.05=α and F_1,3=2.165) indicate no statistically significant difference. The analysis used a 2-factor ANOVA with replication.

On the basis of the results (P value [sessions]=.91>.05=α; F_3,3=0.171), we can conclude that there were no significant differences in the effectiveness of the groups across the different sessions.

There were no significant differences in the interaction between groups and sessions (P value [interactions]=.61>.05=α and F_3,3=0.609).

Of the 5 participants in the non-BF group, 4 played 2 levels using BF. Two of them chose each game version, whereas the other 2 preferred the BF version exclusively.

Overall, BF had a greater effect on motivating patients. With greater efforts to leverage its potential within the game, the positive impact on motivation can be substantially enhanced.

Nevertheless, it is essential to note that the non-BF version fosters motivation by incorporating 2 vital motivational elements: gamification and VR.

As participants enter new and especially challenging stages, their internal motivation to play tends to decrease, whereas their anxiety increases. However, with repeated attempts at this stage, motivation gradually increased, and anxiety levels tended to decrease.

Effect 2: Simulating Fearful Situations

The non-BF group had, on average, 40 points higher anxiety scores across all 4 rounds of the forced games compared with the BF group (the sum of anxiety scores for BF vs non-BF in the first to fourth sessions were: 11 vs 33, 22 vs 34, 29 vs 34, and 27 vs 28). There was a statistically significant difference between the 2 groups (P=.009<.05=α and F_1,3=7.805). The total anxiety score for the non-BF group was 129, whereas that for the other group was 89, indicating the beneficial role of BF in anxiety control. This finding also suggests that using BF could potentially reduce the need for a therapist’s presence. Caution is advised when interpreting these data, as it may be influenced by individuals with severe phobias. The crucial point is that both game variants can evoke anxiety, as they simulate fearful situations. During the games, 5 participants (4 without BF and 1 with BF) experienced extreme stress, necessitating temporary pauses to help them calm down. One participant even reported an increase in blinking frequency when feeling nervous while playing the game.
The $P$ value (sessions)=$0.32 > 0.05 = \infty$ and $F_{1,3}=1.204$, indicating no significant differences in the effectiveness of the groups across different sessions. Many participants experienced anxiety even before the games began, which significantly impacted their anxiety levels during training (picture step). One participant even mistook pictures of cats in the training as real cats because of high tension. In addition, 6 participants (4 without BF and 2 with BF) responded to the cat pictures. On the basis of the data and participant feedback, the order of increasing anxiety levels followed the sequence of stages, starting from the trial game and progressing through the forced games in the following order: fantasy, low-poly, and high-poly cats. Similarly, the normalization of cats occurred in the following order: fantasy cats, pictures of cats, low-poly cats, and high-poly cats. For instance, anxiety levels increased as the number of cats increased. No significant differences in interaction between groups and sessions were observed ($P$ value $[\text{interactions}]=0.20 > 0.05 = \infty$ and $F_{1,3}=1.652$).

**Effect 3: Controlling Stressful Circumstances, Eliminating Therapists’ Presence, and Mitigating Catastrophic Thoughts**

Most participants about the positive signs of reducing their catastrophic thoughts expressed that encountering cats had started to feel somewhat normal. They noted that with continued play, their irrational fears could be replaced with more rational ones, and these positive changes could extend beyond the game to real-life environments. One participant shared, “Before playing the game, I couldn’t even look at cats’ stickers or images, and I used to throw my toy cat out of my room window into the street.” Another participant expressed, “Encountering fantasy cats in small numbers has become normal for me, and I believe that over time, my fear of other types of cats will decrease.” Follow-up data are required to verify the lasting impact of these positive changes.

A total of six noteworthy comments on the elimination of therapists using BF were suggested: (1) after a few sessions, the game can be played independently without therapists; (2) the treatment process can be shortened; (3) patients with milder phobias can benefit from playing without therapists. Otherwise, therapists’ support is necessary during the initial sessions; (4) the game is more beneficial for therapists, offering a controlled environment free of danger; (5) combining virtual and face-to-face treatments is recommended, starting with the game to prepare patients for real-life cat encounters; and (6) BF cannot provide the psychological support therapists offer. One participant, Fatemeh, repeatedly reassured herself during gameplay, saying, “Fatemeh, it’s just a cat, it’s nothing, keep calm.” The necessity of a virtual therapist to provide reassurance and guidance during moments of severe anxiety was evident. Participants either managed to calm themselves or received assistance from us. At times, we had to explain the unlocked stage scenarios to convince the participants to proceed with the remaining games.

To enhance the effectiveness, some participants suggested that the game should display their effort by showing the minimum and maximum HR and the time taken to complete a level. In addition, 2 positive comments regarding HR were as follows: “I noticed that my fear is higher before encountering cats, but my heart rate decreases when I face them” and “Before playing, I believed my fear of cats was overwhelming, but the game helped me realize it wasn’t as intense as I thought.”

**Effect 4: Preliminary Effects on Ailurophobia Treatment**

Using ANOVA single factor, we could not detect a difference between the groups ($F_{1,8}=0.073$, and $P$ value=$0.79 > 0.05 = \infty$). The S-scale scores worsened by 50 and 33 points in the non-BF and BF groups, respectively (Table 2). Both variants induced anxiety, but the BF group showed lower anxiety levels, suggesting that BF was more effective in reducing stress.

No significant difference between the groups was detected ($P$ value=$0.63 > 0.05 = \infty$, and $F_{1,8}=0.256$). The non-BF group improved by 67 points in the FCQ scores, whereas the BF group worsened by 42 points (Table 3).

The significant difference in scores can be attributed to one participant in the non-BF group who initially experienced high anxiety before and during the game. However, as she played more games, her scores on the S-scale (64-28) and FCQ (119-13) decreased dramatically. She mentioned that she used to be greatly bothered by cats being near her or hearing their voices, but after playing the game, she felt less anxious. The constant presence of cats in the game and being able to hear their voices helped her overcome her fears. It is noteworthy that this participant played the game more than all other players, completing 10 levels, including the training stage. In the last 3 stages, the participants specified an anxiety level of 1 out of 10. Initially, we considered this participant’s data as an outlier, but because of the high number of games played, we retained her data. This observation clearly indicates that playing the game more frequently helps to normalize interactions with cats. Her anxiety scores (of 10) for playing 9 levels of the game were the data related to training was excluded for all participants): 10, 10, 8, 3, 3, 2, 1, 1, 1. By replacing her score with a typical number, we obtained more reasonable scores. The non-BF and BF scores worsened by 9 and 42, respectively. Both game versions induced similar anxiety levels in participants. Some of the participants experienced symptom improvement. To assess the initial positive signs of phobia treatment using the FCQ, we should wait until the completion of 10 game stages on average. All participants completed this questionnaire shortly after the games (within a maximum of 10 minutes), and the effects of anxiety caused by fear were still evident. We had to reassure them that the game was not very scary and that the unpredictable event they feared would not happen in the next level, as 4 participants experienced extreme anxiety. These participants took longer breaks between the phases or temporarily stopped playing the game. This anxiety could adversely affect their grades. In addition, approximately 80% (8/10) of the participants mentioned that playing the game more often helped them become accustomed to seeing cats. All participants expressed a preference for the gamified VRET with BF, stating that the experience was more novel and perceived as more effective in reducing fear.
Table 2. Pretest and posttest scores of S-scales.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-BF&lt;sup&gt;a&lt;/sup&gt; (control)</td>
<td>39</td>
<td>61</td>
<td>−22</td>
</tr>
<tr>
<td>Non-BF</td>
<td>64</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>Non-BF</td>
<td>38</td>
<td>67</td>
<td>−29</td>
</tr>
<tr>
<td>Non-BF</td>
<td>46</td>
<td>67</td>
<td>−21</td>
</tr>
<tr>
<td>Non-BF</td>
<td>33</td>
<td>47</td>
<td>−14</td>
</tr>
<tr>
<td>BF (experimental)</td>
<td>39</td>
<td>45</td>
<td>−6</td>
</tr>
<tr>
<td>BF</td>
<td>29</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>BF</td>
<td>42</td>
<td>65</td>
<td>−23</td>
</tr>
<tr>
<td>BF</td>
<td>39</td>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>BF</td>
<td>34</td>
<td>41</td>
<td>−7</td>
</tr>
</tbody>
</table>

<sup>a</sup>BF: biofeedback.

Table 3. Pretest and posttest scores of Fear of Cats Questionnaire.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-BF&lt;sup&gt;a&lt;/sup&gt; (control)</td>
<td>71</td>
<td>100</td>
<td>−29</td>
</tr>
<tr>
<td>Non-BF</td>
<td>119</td>
<td>13</td>
<td>106</td>
</tr>
<tr>
<td>Non-BF</td>
<td>82</td>
<td>94</td>
<td>−12</td>
</tr>
<tr>
<td>Non-BF</td>
<td>112</td>
<td>92</td>
<td>20</td>
</tr>
<tr>
<td>Non-BF</td>
<td>101</td>
<td>119</td>
<td>−18</td>
</tr>
<tr>
<td>BF (experimental)</td>
<td>84</td>
<td>97</td>
<td>−13</td>
</tr>
<tr>
<td>BF</td>
<td>70</td>
<td>82</td>
<td>−12</td>
</tr>
<tr>
<td>BF</td>
<td>76</td>
<td>104</td>
<td>−28</td>
</tr>
<tr>
<td>BF</td>
<td>89</td>
<td>85</td>
<td>4</td>
</tr>
<tr>
<td>BF</td>
<td>89</td>
<td>82</td>
<td>7</td>
</tr>
</tbody>
</table>

<sup>a</sup>BF: biofeedback.

Patients’ Preferences About the Designed Treatment

Most participants expressed that the game had a positive therapeutic impact and was capable of normalizing their interactions with cats. In total, 2 participants played the game 7 and 10 times and reported significant changes in their perception of cat-related fears. They shared that their perceptions of cat fear transformed, and encountering cats felt normal. Moreover, they believed that this effect could be extended to real-life situations. One participant shared, “I used to feel uneasy when cats were nearby, and the sound of cats was distressing for me. But now, as cats are consistently present in the game, and the sound of cats is played during the gameplay, being close to cats and hearing their sounds has become completely normal for me.” In the last 2 stages, their anxiety levels were reduced to a rating of 1 of 10. Before playing, most participants anticipated that cats would appear in the forest and perch on the tree branches. They expected the paths where cats were located to have denser and more crowded areas, featuring an abundance of trees, wooden huts, and gazebos with cats nearby. One commented, “The space provided is too vast, and it could be made more intense to induce fear. It would be beneficial to create some narrower paths leading to a door where cats are positioned. This could instill more fear. Generally, the game’s paths are not challenging situations, and a darker environment could make the cat’s eyes more prominent.” These comments contradict most participants, who appreciated the game’s positive aspect of indirectly implementing treatment and displaying everyday interactions people have with cats. Incorporating various environments and cat behaviors could further normalize the interaction with cats from all angles. However, these changes must be introduced with caution to avoid reinforcing the perception that cats are scary. In addition, the suggested locations to be included in the game range from the park environment to urban settings, such as apartments, streets, alleys, markets, cafés, dark scenes, kitchens, and garbage cans.

Some participants preferred the fantasy cats, believing that they alone have the ability to normalize interactions with cats because they highlight the positive aspects of cats such as their beautiful eyes and portray them as attractive, safe, and less harmful. Designing different fantasy cats appears to be a reasonable way to encourage individuals. One participant said, “It bothers me...
that the cats’ heads are small and their tails are long. In contrast, fantasy models had big heads and short tails. In different game levels, placing fantasy cats next to other cats conveys the feeling that all cats are harmless. Starting with images of rough and fat cats and gradually increasing the number of cats, and transitioning them to real models, helped me realize that the initial stage’s image was merely in my mind and unreal. As the cats’ numbers increased, I discovered that they did not pose any harm.” These eye-opening opinions shed light on an overlooked aspect—the psychological impact of the game’s difficulty levels and cat types.

Preferring fantasy cats indirectly revealed that low- and high-poly cats mostly evoke fear. Most participants found these cats to be more rational. Increased aversion and avoidance were observed in places with more cat voices and presence. Longer sounds also intensified fear.

On the basis of these findings, it is suggested to gradually introduce sounds, starting from cats with no sounds to short and pleasant sounds and then to real single and multiple sounds. The maximum fear was near the boxes where the number and noise of the cats were higher. Although this arrangement was found to be effective and logical in normalizing interactions with cats, high fear levels may have led some participants to avoid playing altogether. One participant preferred orderly and group cats for a calmer experience, whereas disorderly placement near the box increased fear. These reasons highlight the significance of using fantasy cats. Most participants found the size of the cats were found to be suitable. However, larger cat sizes, such as pictures of striped cats and low-poly cats, increased anxiety. The picture level, considered the easiest, induced anxiety and fear in most participants (6 of 10). Concerning cats’ behavior, most preferred nonreactive cats, such as fantasy cats that simply look at the sky in a cartoony manner; cats sitting and grooming themselves; or cats moving along the path without any reaction. Most participants disliked black cats waving their hands or white cream cats turning and staring at the player.

Most participants expressed the need for the game’s cat designs to closely resemble real-world cats. The following cats were not used based on their comments:

- Spotted (mainly black and white) and gray-striped cats, which are abundant in Iran.
- Kittens: Participants made three points: (1) kittens may not have a significant therapeutic effect, but they enhance the game’s appeal and create a more lifelike environment; (2) the treasure finder can be replaced with a fantasy kitten, allowing for a more captivating display of less favored features of cats, such as their nails, tail, and head. Moreover, their beautiful eyes can be showcased as larger; and (3) the option of raising a kitten in the game.
- Fierce-looking cats with grabbing capabilities: adding them requires expert opinions. Although statistics on cat grabbing are limited, the actual occurrence is likely to be minimal. People’s intense fears may exaggerate this concern.
- A mother cat breastfeeding her babies for a heartwarming and motherly touch.
- Sphynx cats: despite being rare in Iran, could enhance realism and normalize fear of diverse cat breeds.
- Fat or fluffy cats resembling a doll-like appearance.
- Placing cats amidst the greens and bushes along the paths.
- Injured (eg, cats with one eye or leg) or lifeless cats.
- Sudden movements of cats (eg, cats leaping out of trash cans): mentioned by most participants.
- Feeding cats: some participants did not agree with implementing this feature.

In conclusion, the game layout and models were considered logical by most of the participants. They stressed that fighting with cats in the game could worsen their fear, making it a clear distinction between a therapeutic game and one designed solely for entertainment purposes. This opinion is in agreement with the clinical expert (the fourth author) who emphasized that the games for treating animal phobias should avoid action and fighting scenarios. For example, reducing the fear of cockroaches using scenarios where they stomp on or kill them may adversely affect.

**Heuristics Evaluations**

As presented in Tables 4-7, of the 44 heuristics adapted from the Nielsen user interface, VR, and playability, an impressive 41 principles obtained scores of 62% or higher, underscoring the tool’s potential as a therapeutic product. Moreover, it enhances patient adherence to the treatment process.

Overall, 90% (9/10) of the participants found learning to play the game remarkably easy, particularly with the convenience of using just one button under VR glasses, which proved beneficial for those with mobility disabilities. Two suggestions emerged concerning in-game movement: (1) incorporating a back button and (2) movement through walking, potentially achieved with motion-sensing devices. However, careful consideration is necessary to ensure that they positively impact the player experience. Some individuals may prefer a less cumbersome setup. To enhance experience, it is crucial to incorporate a tutorial in a video or audio format for first-time users by introducing relaxation techniques to manage panic situations. Many participants required clarification that frightful situations would not occur at the subsequent levels. Providing detailed descriptions of new levels, including information about cats’ types and behaviors, prevents players from creating self-made stories about cat attacks. Moreover, to improve clarity, players needed clearer instructions after opening each box, signaling that they should open 4 boxes per session. Although a ribbon in the corner of the screen displays the number of opened boxes, it does not adequately alert the players to this requirement. Among the VR principles, the navigation and orientation support principle excelled at 82%, with patients being well-informed about their in-game position. Notably, approximately 80% (8/10) of patients experienced no dizziness during extended gameplay. To increase the level of engagement and therapeutic impact, introducing a punishment mechanism, such as reducing players’ points, could be beneficial. It might be worth reconsidering the features of allowing players to win the game without encountering cats. Game sounds and music received a relatively low score (51%), causing tension and unease, instead of promoting peace and happiness. The addition
of soothing natural sounds was also suggested. In addition, consider a sound to indicate proximity to the box, reducing the need to check the bar constantly and improving the focus on gameplay. The game could benefit from a save and resume feature, especially during panic situations, allowing patients to take a moment to calm down. Some also raised concerns about the suitability of graphics for older adult audiences.

The principle of variety in the game’s paths and challenges stands out as one of the main gameplay principles. Although it obtained a relatively good score (68%), most participants said that after a few stages, the game became monotonous. Players quickly realized that cats only appear in certain sections of the roads and near treasure boxes. Certain adjustments were recommended to enhance the game’s appeal. Increasing the spacing between trees and raising their height can create a more immersive environment. Adding colorful elements such as flowers, toys, water views, and a gazebo in the park will infuse vibrancy into the game. In general, elevating the game’s attractiveness can be achieved by introducing a greater sense of adventure without relying on unrealistic fears. One participant suggested that instead of having the treasure box as the game’s goal, it could be placed in various locations within the forest, each rewarding the player with different prizes, such as food. Another suggestion was to replace the guide bar, which received positive feedback from the participants, with a map that indicated the approximate distance to the target. In addition, the introduction of a captivating and fantastical cat character instead of the current bar was recommended. In total, 2 participants pointed out that displaying HR in the corner might be somewhat distracting. It was suggested to show HR only when it was high or to remind players to reduce stress using a heartbeat’s sound.
<table>
<thead>
<tr>
<th>Usability heuristic and question</th>
<th>Question, mean (SD)</th>
<th>Heuristic, mean (SD)</th>
<th>Heuristic overall percent, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Natural engagement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>3.5 (1.08)</td>
<td>3.7 (0.28)</td>
<td>74</td>
</tr>
<tr>
<td>Q2</td>
<td>3.9 (0.88)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>2. Compatibility with the user's task and domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>3.7 (0.95)</td>
<td>3.7 (0.3)</td>
<td>74</td>
</tr>
<tr>
<td>Q4</td>
<td>4 (1.25)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Q5</td>
<td>3.4 (1.17)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>3. Natural expression of action</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>3.5 (1.27)</td>
<td>3.2 (0.42)</td>
<td>64</td>
</tr>
<tr>
<td>Q7</td>
<td>2.9 (1.20)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>4. Close coordination of action and representation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>3.5 (1.27)</td>
<td>3.6 (0.1)</td>
<td>72</td>
</tr>
<tr>
<td>Q9</td>
<td>3.6 (1.17)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Q10</td>
<td>3.7 (1.25)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>5. Realistic feedback</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q11</td>
<td>3.7 (1.4)</td>
<td>3.7 (1.4)</td>
<td>74</td>
</tr>
<tr>
<td><strong>6. Faithful viewpoints</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>3.6 (1.2)</td>
<td>3.6 (1.2)</td>
<td>72</td>
</tr>
<tr>
<td><strong>7. Navigation and orientation support</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q13</td>
<td>4.1 (1.2)</td>
<td>4.1 (1.2)</td>
<td>82</td>
</tr>
<tr>
<td><strong>8. Visibility of system status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q14</td>
<td>3.7 (1.25)</td>
<td>3.63 (0.75)</td>
<td>72.5</td>
</tr>
<tr>
<td>Q15</td>
<td>3.4 (1.17)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Q16</td>
<td>3.7 (1.49)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Q17</td>
<td>3.7 (1.16)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>9. Consistency and standards</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q18</td>
<td>3.7 (1.3)</td>
<td>3.7 (1.3)</td>
<td>74</td>
</tr>
<tr>
<td><strong>10. Error prevention</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q19</td>
<td>3.7 (1.34)</td>
<td>3.5 (0.28)</td>
<td>70</td>
</tr>
<tr>
<td>Q20</td>
<td>3.3 (2)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>11. Recognition rather than recall</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q21</td>
<td>3.6 (1.2)</td>
<td>3.6 (1.2)</td>
<td>72</td>
</tr>
<tr>
<td><strong>12. Flexibility and efficiency of use</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q22</td>
<td>3.8 (1.0)</td>
<td>3.8 (1.0)</td>
<td>76</td>
</tr>
<tr>
<td><strong>14. Help and documentation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q23</td>
<td>2.9 (1.4)</td>
<td>2.9 (1.4)</td>
<td>58</td>
</tr>
</tbody>
</table>

*N/A: not applicable.*
Table 5. Results of the questionnaire designed based on gameplay part of the playability heuristics [52].

<table>
<thead>
<tr>
<th>Question</th>
<th>Heuristic overall percent</th>
<th>Heuristic, mean (SD)</th>
<th>Question, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Player’s fatigue is minimized by varying activities and pacing during game play.</td>
<td>68</td>
<td>Q1: 3.4 (1.6)</td>
<td>3.4 (1.6)</td>
</tr>
<tr>
<td>2. Provide consistency between the game elements and the overarching setting and story to suspend disbelief.</td>
<td>68</td>
<td>Q2: 3.4 (1.3)</td>
<td>3.4 (1.3)</td>
</tr>
<tr>
<td>3. Provide clear goals, present overriding goal early as well as short-term goals throughout play.</td>
<td>86</td>
<td>Q3: 4.3 (1.1)</td>
<td>4.3 (1.1)</td>
</tr>
<tr>
<td>4. There is an interesting and absorbing tutorial that mimics gameplay.</td>
<td>90</td>
<td>Q4: 4.5 (1.0)</td>
<td>4.1 (0.57)</td>
</tr>
<tr>
<td>5. The game is enjoyable to replay.</td>
<td>N/A</td>
<td>Q5: N/A</td>
<td>3.7 (0.9)</td>
</tr>
<tr>
<td>6. Game play should be balanced with multiple ways to win.</td>
<td>70</td>
<td>Q6: 3.5 (0.7)</td>
<td>3.5 (0.7)</td>
</tr>
<tr>
<td>7. Player is taught skills early that you expect the players to use later, or right before the new skill is needed.</td>
<td>68</td>
<td>Q7: 3.8 (1.0)</td>
<td>3.8 (1.0)</td>
</tr>
<tr>
<td>8. Players discover the story as part of gameplay.</td>
<td>80</td>
<td>Q8: 3.4 (1.5)</td>
<td>3.4 (1.5)</td>
</tr>
<tr>
<td>9. The game is fun for the player first, the designer second and the computer third. That is, if the nonexpert player’s experience is not put first, excellent game mechanics and graphics programming triumphs are meaningless.</td>
<td>78</td>
<td>Q9: 4.0 (0.8)</td>
<td>4.0 (0.8)</td>
</tr>
<tr>
<td>10. Player should not experience being penalized repetitively for the same failure.</td>
<td>86</td>
<td>Q10: 3.9 (1.2)</td>
<td>3.9 (1.2)</td>
</tr>
<tr>
<td>11. Player’s should perceive a sense of control and impact onto the game world. The game world reacts to the player and remembers their passage through it. Changes the player makes in the game world are persistent and noticeable if they back-track to where they have been before.</td>
<td>82</td>
<td>Q11: 4.3 (0.7)</td>
<td>4.3 (0.7)</td>
</tr>
<tr>
<td>12. The game should give rewards that immerse the player more deeply in the game by increasing their capabilities (power-up), and expanding their ability to customize.</td>
<td>70</td>
<td>Q12: 4.1 (1.0)</td>
<td>3.9 (0.28)</td>
</tr>
<tr>
<td>13. Pace the game to apply pressure but not frustrate the player. Vary the difficulty level so that the player has greater challenge as they develop mastery. Easy to learn, hard to master.</td>
<td>78</td>
<td>Q13: 3.7 (1.3)</td>
<td>N/A</td>
</tr>
<tr>
<td>14. Challenges are positive game experiences, rather than a negative experience (results in their wanting to play more, rather than quitting).</td>
<td>76</td>
<td>Q14: 3.5 (1.3)</td>
<td>3.7 (0.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q15: 3.9 (1.0)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q16: 3.9 (1.2)</td>
<td>3.75 (0.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q17: 3.6 (1.2)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q18: 3.8 (1.1)</td>
<td>3.8 (1.1)</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
Table 6. Results of the questionnaire designed based on mechanic part of the playability heuristics [52].

<table>
<thead>
<tr>
<th>Question</th>
<th>Question, mean (SD)</th>
<th>Heuristic, mean (SD)</th>
<th>Heuristic overall percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Game should react in a consistent, challenging, and exciting way to the player’s actions (eg, appropriate music with the action).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>2.8 (1.6)</td>
<td>2.8 (1.6)</td>
<td>51</td>
</tr>
<tr>
<td>2. Make effects of the AI(^a) clearly visible to the player by ensuring they are consistent with the player’s reasonable expectations of the AI actor.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>3.1 (0.9)</td>
<td>3.1 (0.9)</td>
<td>62.3</td>
</tr>
<tr>
<td>3. A player should always be able to identify their score/status and goal in the game.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>4.4 (0.5)</td>
<td>4.3 (0.14)</td>
<td>86</td>
</tr>
<tr>
<td>Q4</td>
<td>4.2 (0.9)</td>
<td>N/A(^b)</td>
<td></td>
</tr>
<tr>
<td>4. Mechanics/controller actions have consistently mapped and learnable responses.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>4.4 (1.1)</td>
<td>4.15 (0.35)</td>
<td>83</td>
</tr>
<tr>
<td>Q6</td>
<td>3.9 (1.6)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>5. Shorten the learning curve by following the trends set by the gaming industry to meet user’s expectations.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q7</td>
<td>4.3 (1.3)</td>
<td>3.95 (0.49)</td>
<td>79</td>
</tr>
<tr>
<td>Q8</td>
<td>3.6 (1.6)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>6. Controls should be intuitive, and mapped in a natural way; they should be customizable and default to industry standard settings.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q9</td>
<td>4.5 (1.0)</td>
<td>4.35 (0.21)</td>
<td>87</td>
</tr>
<tr>
<td>Q10</td>
<td>4.2 (0.9)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>7. Player should be given controls that are basic enough to learn quickly yet expandable for advanced options.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q11</td>
<td>3.3 (1.3)</td>
<td>3.53 (0.32)</td>
<td>70.67</td>
</tr>
<tr>
<td>Q12</td>
<td>3.4 (1.7)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Q13</td>
<td>3.9 (1.4)</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)AI: artificial intelligence.

\(^b\)N/A: not applicable.
Table 7. Results of the questionnaire designed based on usability part of the playability heuristics [52].

<table>
<thead>
<tr>
<th>Question</th>
<th>Question, mean (SD)</th>
<th>Heuristic, mean (SD)</th>
<th>Heuristic overall percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Provide immediate feedback for user actions.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>4.1 (1.5)</td>
<td>4.1 (1.5)</td>
<td>82</td>
</tr>
<tr>
<td>2. The player can easily turn the game off and on, and be able to save games in different states.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>2.3 (1.3)</td>
<td>2.3 (1.3)</td>
<td>46</td>
</tr>
<tr>
<td>3. The player experiences the user interface as consistent (in control, color, typography, and dialog design) but the gameplay is varied.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>3.7 (1.3)</td>
<td>3.35 (0.49)</td>
<td>67</td>
</tr>
<tr>
<td>Q4</td>
<td>3 (1.2)</td>
<td>N/A (^a)</td>
<td>N/A</td>
</tr>
<tr>
<td>4. The player should experience the menu as a part of the game.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>3.4 (1.0)</td>
<td>3.65 (0.35)</td>
<td>68</td>
</tr>
<tr>
<td>Q6</td>
<td>3.9 (1.0)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>5. Sounds from the game provide meaningful feedback or stir a particular emotion.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q7</td>
<td>3.5 (1.2)</td>
<td>3.35 (0.35)</td>
<td>67</td>
</tr>
<tr>
<td>Q8</td>
<td>3.2 (1.1)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Q9</td>
<td>2.8 (1.6)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>6. Players do not need to use a manual to play the game.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>4 (0.9)</td>
<td>4 (0.9)</td>
<td>80</td>
</tr>
<tr>
<td>7. Make the menu layers well organized and minimalist to the extent the menu options are intuitive.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q11</td>
<td>3.9 (1.6)</td>
<td>3.9 (1.6)</td>
<td>78</td>
</tr>
<tr>
<td>8. Get the player involved quickly and easily with tutorials and/or progressive or adjustable difficulty levels.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>4.1 (1.3)</td>
<td>3.95 (0.21)</td>
<td>79</td>
</tr>
<tr>
<td>Q13</td>
<td>3.8 (1.4)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>9. Art should be recognizable to the player, and speak to its function.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q14</td>
<td>3.7 (1.2)</td>
<td>3.65 (0.07)</td>
<td>73</td>
</tr>
<tr>
<td>Q15</td>
<td>3.6 (0.7)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\(^a\)N/A: not applicable.

**Discussion**

**Principal Findings**

We developed a gamified VRET augmented with BF to address ailurophobia. To our knowledge, no specialized research on ailurophobia treatment exists, either in Iran or internationally. Motivated by the high prevalence of ailurophobia and the lack of accessible gamified VR environments with BF, our main goal was to create and assess a smartphone-based VRET augmented with BF for animal phobia (cat phobia). We hypothesized that this tool would better motivate patients, manage stress, simulate fearful situations, treat phobia, and reduce therapists’ involvement compared with a gamified VRET. The tool was designed based on expert sessions in video games, gamification, cognitive, and psychology. The results indicate its positive impact on specified features. Of the 44 heuristics, 41 scored above 62%, showing the potential for phobia interventions and motivating patients for treatment. Although tested on only 10 participants for a short duration (up to 3 hours without follow-up sessions), the results were reliable. Extensive data and feedback collection have been used to evaluate various aspects of the tool. On average, after 10 sessions, initial signs of improvement were observed, with slight variations depending on individuals’ phobia levels. One intriguing finding was that most participants were content with the game’s indirect approach to normalize interaction with cats and its nonviolent nature. They emphasized that action or combat scenarios would reinforce unrealistic fears and validate their phobia. Another significant finding was the progression of normalization in dealing with cats, tolerating their behavior, and hearing their voices, which gradually became more challenging. Although the current game normalizes communication with cats and holds good appeal, most participants suggested improvements, such as adding a variety of cats that closely resemble real-world characteristics, including voices and behaviors, to further enhance the normalization process. In addition, most participants expressed satisfaction with the game’s easy movements and minimal learning curve. To enhance the experience, adding diversity and adventure while minimizing unrealistic violence was recommended. Moreover, during the evaluations, the participants strongly felt the need for a virtual therapist to provide calming guidance and support during moments of severe anxiety.
Comparison With Prior Work

To our knowledge, no study has simultaneously used BF, VR, and gamification for the treatment of animal phobia. However, various studies have used VR and game concepts to address specific animal phobias, for example, spider phobia [12,32] and snake phobia [47]. Similar to these studies, our tool successfully induced anxiety and led to a reduction in fear levels, avoidance behaviors, and catastrophic thoughts related to phobias. In addition, it positively boosted their motivation for treatment adherence. Unlike previous studies, our unique feature was the initial evaluation, showing that participants preferred a gamified VRET with BF. It has proven to be more effective in reducing symptoms and increasing internal motivation. These findings align with recent reviews highlighting the significant anxiety-reducing benefits of combining VR and BF, along with the advantages in motivation, user experience, involvement, and attentional focus [53,54]. In contrast to our study, where more participants preferred interacting with safe stimuli such as fantasy cats, studies such as those by Dibbets and Schruers [55] and Pittig et al [56] reported that selecting riskier options led to a stronger decrease in self-reported spider fear and disgust, whereas safe choices increased these emotions. The differing outcomes could be attributed to the use of VR and 3D images. VRETs are widely recognized as an appealing treatment modality because of their perceived naturalness in the automated format. However, Albakri et al [57] suggested that augmented reality exposure therapies offer a better experience and increased realism by seamlessly integrating digital information into the real world rather than creating a completely new virtual environment. We plan to explore the implementation of our designed tool with augmented reality and compare the outcomes in future studies. Dibbets and Schruers [55] found that the number of spiders encountered did not correlate with declines in aversive feelings and avoidance behaviors. However, our study concluded that a higher number of stimuli were more effective in normalizing interactions with cats. In addition, we observed that the action and combat scenarios were not beneficial for individuals with phobias. Interestingly, snake phobia treatment in a nearly action genre format [47] lacks a rationale for its selection. Further research is required to determine and devise appropriate scenarios for individuals with phobias. Throughout this study, the need to conduct similar research was highlighted. It was not feasible to make precise comparisons with prior studies in every detail.

Limitations

The initial study on treating ailurophobia using VRETs with gamification and BF had limitations, primarily a small number of respondents. A total of 10 potential participants were inaccessible after Instagram’s temporary filtering in our country. The sample was skewed toward educated participants in their twenties and thirties, indicating the need to include diverse educational backgrounds, children, adolescents, and older adults. Owing to time constraints, we did not use any statistical method to calculate the required sample size. The study by Mor et al [48] recommended a minimum of 20 participants in each study arm for feasibility pilot studies on treating flying phobia using 360° images. Certainly, a larger number of patients is needed in each arm for the primary assessments. One future work is to replicate the study quantitatively and more rigorously while also introducing another arm that uses standard and clinical exposure therapies, enabling us to evaluate the tool and showcase more applications. In addition, the small sample sizes prohibited us from examining dropout rates. The results are exploratory, and long-term effects remain unknown due to the lack of follow-up data. Only one self-rating scale, the FCQ, has been used to diagnose individuals with ailurophobia. However, it is advisable to supplement such questionnaires with a telephone or face-to-face diagnostic interview conducted by an expert clinician, typically lasting approximately 30 minutes [2,12,32]. These interviews not only boost diagnostic reliability but also enable descriptive analysis [2]. It is worth mentioning that the participants were initially asked to explain the origin and signs of their ailurophobia. Participants were randomly divided into groups; however, the equality of their stress levels was not considered. It appears that by preserving randomness, the stress levels of individuals in the study groups should be nearly equal. For example, if one group has 2 extreme cases, the other groups should also have 2 similar cases to ensure transparency and enhance the reliability of the results. Creating a real-world game proved challenging owing to the limitations of the smartphone platform. Although playability and system usability questionnaires were not rigorously assessed, they were designed based on popular usability scales, including Nielsen [49,50], VR [51], and playability [52] heuristic evaluations. Changes in the individual’s physiological status, particularly HR, influence their experience. Unfortunately, this feature could not be assessed in the BF arm owing to the small sample size. Understanding its effectiveness in high-tension situations and its role in reducing anxiety remains a top priority.

Conclusions and Future Work

The gamified VRET incorporating BF for treating cat phobia could be effective and has the potential to evolve into a comprehensive tool. One way to enhance its utility is by expanding the variety of cat types and behaviors, simulating different environments where cats are commonly found, and boosting its appeal through increased adventure while avoiding the use of unrealistic fears. After modifying the tool and using more robust study designs with ample sample sizes, further investigation can explore how this tool can be used in treatments without the presence of a therapist or combined (virtual and real simulation of fear), both in clinics and remotely. The park environment has the potential to effectively treat various animal phobias and other specific phobias. Implementing a gradual progression of sound stimuli could improve the therapeutic process. Starting with serene and pleasant sounds and gradually advancing to more challenging and potentially distressing voices, like cats squealing (inspired by a participant’s recollection of hearing a cat giving birth) or their aggressive vocalizations during fights. The final suggestion is to add the possibility of interacting with cats during more challenging stages, thereby bridging the game environment with the real world.
Acknowledgments
All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Data Availability
The data presented in this study are available from the corresponding author upon request.

Authors' Contributions
AK contributed to the conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing the original draft, and reviewing and editing. AN contributed to conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing the original draft, and reviewing and editing. ZA contributed to the conceptualization, data curation, formal analysis, methodology, project administration, resources, validation, visualization, writing the original draft, and reviewing and editing. AKB and PHA contributed to conceptualization, formal analysis, funding acquisition, investigation, resources, software, validation, and reviewing and editing.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Questionnaires were designed based on user interfaces, virtual reality, and playability heuristics, along with their results. [DOC File, 136 KB - games_v12i1e34535_app1.doc ]

References


40. Mahram B. Standardization of spilberger anxiety test in Mashhad. Allameh Tabataba’i University. 1994. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=ICB0EFwAAAAJ&citation_for_view=ICB0EFwAAAAJ:Ti5es2fBq6C [accessed 2024-01-01]


Abbreviations

**BF:** biofeedback  
**FCQ:** Fear of Cats Questionnaire  
**FSQ:** Fear of Spiders Questionnaire  
**HR:** heart rate  
**STAI:** State-Trait Anxiety Inventory  
**VR:** virtual reality  
**VRET:** virtual reality exposure therapy
An Augmented Reality Serious Game for Children’s Optical Science Education: Randomized Controlled Trial

Bo Liu, MA; Xinyue Wan, MA; Xiaofang Li, MA; Dian Zhu, PhD; Zhao Liu, PhD
School of Design, Shanghai Jiao Tong University, Shanghai, China

Corresponding Author:
Zhao Liu, PhD

Abstract

Background: Knowledge construction in the context of children’s science education is an important part of fostering the development of early scientific literacy. Nevertheless, children sometimes struggle to comprehend scientific knowledge due to the presence of abstract notions.

Objective: This study aimed to evaluate the efficacy of augmented reality (AR) games as a teaching tool for enhancing children’s understanding of optical science education.

Methods: A total of 36 healthy Chinese children aged 6-8 years were included in this study. The children were randomly divided into an intervention group (n=18, 50%) and a control group (n=18, 50%). The intervention group received 20 minutes of AR science education using 3 game-based learning modules, whereas the control group was asked to learn the same knowledge for 20 minutes with a non-AR science learning app. Predict observe explain tests for 3 topics (animal vision, light transmission, and color-light mixing) were conducted for all participants before and after the experiment. Additionally, the Intrinsic Motivation Inventory, which measures levels of interest-enjoyment, perceived competence, effort-importance, and tension-pressure, was conducted for children after the experiment.

Results: There was a statistically significant difference in light transmission ($z=-2.696; P=0.008$), color-light mixing ($z=-2.508; P=0.01$), and total predict observe explain test scores ($z=2.458; P=0.01$) between the 2 groups. There were also variations between the groups in terms of levels of interest-enjoyment ($z=-2.440; P=0.02$) and perceived competence ($z=-2.170; P=0.03$) as measured by the Intrinsic Motivation Inventory.

Conclusions: The randomized controlled trial confirmed that the AR-based science education game we designed can correct children’s misconceptions about science and enhance the effectiveness of science education.

Trial Registration: ClinicalTrials.gov NCT06184022; https://classic.clinicaltrials.gov/ct2/show/NCT06184022

(JMIR Serious Games 2024;12:e47807) doi:10.2196/47807

KEYWORDS
augmented reality; serious game; science education; childhood education; cognition; children; scientific cognition; cognitive process; effectiveness

Introduction

Children’s level of scientific concept generation is representative of their inquiry, comprehension, and application of natural events and phenomena and reveals their cognitive capacities and developmental stages [1]. Knowledge construction in children’s science education contributes to early scientific literacy development, which improves children’s cognitive level by enhancing thinking skills, and is being emphasized by scholars and parents [2]. Traditionally, children build domain knowledge in science through films, literature, and lectures in science education [3]. Although some forms of educational learning are accessible, they often use a monotonous instructional format, and confusing content hinders the transmission of scientific knowledge [4].

Serious games provide a more engaging interactive environment and an accessible cognitive framework to facilitate effective learning [5]. Studies have shown that serious games have more effective learning outcomes than traditional methods of science education (eg, face-to-face lectures and book-based knowledge transfer) [6,7]. It is suitable for children’s investigation of natural phenomena because the game’s visual design simulates paranormal phenomena that cannot be produced in real life. Lester et al [8] constructed virtual environments that generate natural phenomena, allowing children to assume roles in open-world environments and to freely rely on their knowledge of the geography and biology of natural environments. Laine et al [9] permitted children to interact with hosts in virtual narrative game scenarios and to investigate the geometry of the virtual environment with the protagonist. The concept of light, a prevalent natural phenomenon, was selected as the subject of this research to explore its design for enhancing children’s
cognitive abilities. Optical science education programs are still presented in a 2D format, which has been demonstrated to be ineffective [10].

Due to the spatial complexity and abstract nature of optics, it is challenging to accurately convey knowledge through flat visual representations [11]. Therefore, it is necessary to blur the boundaries between the 3D real world and the 2D digital world to reduce the distance between children’s learning of science concepts and their learning environments [12]. 3D representations and interactions in augmented reality (AR) games have the potential to enhance spatial cognition, thereby facilitating children’s comprehension of spatially abstract scientific concepts [13], such as simulating the movement of the sun in a classroom environment [14]. Sahin and Yilmaz [15] demonstrated that students who used AR technology to improve their science literacy performed better on tests than those who learned using traditional methods. This is as a result of AR technology’s ability to enhance the dynamic potential of human consciousness to comprehend the science learning process [16]. In addition, motivational improvement was mentioned as one of the frequently observed AR outcomes [17]. Using AR apps increased student motivation relative to other instructional aids [18]. Our study investigated whether designing optical science education with more comprehensible 3D interactions for children can enhance science education and promote children’s motivation.

The study designed the “AR Serious Game for Optical Science” and conducted a randomized controlled trial to determine the efficacy of this AR game product in enhancing children’s science education. The primary objective of this study was to validate the efficacy of AR science education games for children; the secondary objective was to investigate the intrinsic motivation of children toward them.

**Methods**

**Study Design**

Guardians of children with independent mobility provided informed written consent for their participation in the study. Participants were randomly assigned to the intervention and control groups using a randomization list, which was maintained by members of the study group uninvolved in any other aspect of the study. Participants’ guardians received and opened opaque, sealed envelopes containing group assignments following the initial evaluation. The evaluator in charge of assessing the results of the AR science education course had no access to participant information or group assignment.

Sample size calculations were performed using PASS software (NCSS LLC) based on the predict observe explain (POE) test scores from the preintervention questions. Group sample sizes of 18 and 18 achieve 90.118% power to reject the null hypothesis of equal means, when the population mean difference is \( \mu_1 - \mu_2 = 3.2 - 1.0 = 2.2 \), with SDs of 2.0 and 1.9 for the 2 groups and with a significance level (\( \alpha \)) of .05 using a bilateral, 2-sample, equal-variance, 2-tailed \( t \) test.

**Participants**

A total of 36 Chinese children (aged 6-8 y) were recruited from Jiangyin Children’s Education Center and Jiangyin Wuxi Community in Jiangsu Province and divided into the intervention (n=18, 50%) and control (n=18, 50%) groups.

**AR Science Education Game Design**

During the learning phase, children are required to engage in physical activities, such as walking around with a handheld device, to interact with the AR scene’s content to discover what is unique about the light phenomenon. When children touch the interactive points, the content is explained by animation and voice-overs. This study developed several interactive approaches for children within AR games, such as through in-game visual representations, speech, and interactive methods, which permit children to connect game content to unfamiliar information as they explore. The advantages are as follows: (1) children can use more familiar physical activities with light concepts to establish metaphorical mappings related to orientation, not just gestural touch; (2) rendering light with 3D attributes in the real world reduces the cognitive load generated by children’s linkage of abstract knowledge and the real phenomena; and (3) adding various kinds of digital augmentation effects in the AR scene helps children understand the concepts. The project created 3 games based on the characteristics of scientific understanding (Multimedia Appendix 1 and Figure S1 in Multimedia Appendix 2 [9,14,19-28]).

Game 1 introduces children to the fundamentals of animal vision (Figure 1). Animal vision concepts are investigated through AR scenes. By clicking on the icons in the lower-left corner, the game transforms to an animal simulation. In each scenario, a voice-over narration instructs children to identify the visual differences between the animal and the human. When the handheld device is trained on a specific target, a voice-over narration and feedback animation will play.
In the design of the interaction mode, 3 display modes were established for the game’s interactive elements: far, medium, and near (hybrid camera mode). The concept of invisible light is introduced to children in greater detail based on the ray distance between the device’s camera and the target element. The far view provides children with an intuitive impression of the invisible light’s overall effect; the medium view uses transition animation to illustrate the invisible light’s characteristics; and the near view uses special effect particles to illustrate the invisible light’s trajectory.

Game 2 introduces children to light transmission-related concepts (Figure 2). In the AR scenario, children navigate the environment with a handheld device and activate energy panels by interacting with flat mirrors and optics. By targeting AR-enhanced prop objects and manipulating the angle of light emission to investigate how light propagates, voice-over explanations and feedback animations are activated.
According to the voice-over prompts, children can hold the device and manipulate the flashlight from a first-person perspective (spatial exploration mode) as part of the interactive design. They then complete 3 steps: locating the interactive elements (mirrors, ice crystals, etc), adjusting the flashlight’s tilt angle, and using the flashlight to complete the light-up task. The progression encourages children to investigate the principles of light transmission through the game.

Game 3 introduces children to color-light mixing concepts (Figure 3). Children were instructed to walk around with the device in hand and explore the color changes of props such as AR-enhanced birds, which are illuminated with various colors of lights. Collecting the target color’s shadow initiates a voice-over explanation and feedback animation.
Regarding the interactive design, children need to hold the device to illuminate the creatures and cast shadows on the present wall, and then they need to press the button to turn the light on and off (projection irradiation mode). The objective of the game required children to perform single-color illumination, 2-color mixing, and 3-color mixing to achieve the desired hue. In another vibrant nursery game, children were instructed to move plants to receive various colors of light and to observe the plants’ root elongation and leaf dispersal.

This game design used Unity 3D (Unity Technologies) as the development engine, and the app was installed on an Apple iPad (2018) with a screen resolution of 2048×1536 (264 pixels per inch). The AR component made use of the Vuforia AR SDK (Parametric Technology Corporation) to accomplish the fundamental duties of plane identification and virtual object generation. The interaction section used lens focus to determine the interactions; when the device camera’s output rays collide with the target virtual object and the distance is close, it is deemed to have located the target effectively. To imitate the illusion of invisible light, Unity’s Post Processing module was applied to the camera filter. The principle entailed presenting the camera screen into the buffer of Unity and applying filters and effects prior to displaying it; it can be applied to both the camera screen and the virtual item.

**Procedure**

This experiment was a randomized controlled trial, and the participants were randomly separated into the intervention group.
and the control group. The random numbers were generated by applying the SAS software analysis system (SAS Institute) on a computer simulation, and no experimental group was allowed to be selected at random. Every child was tested in the company of a guardian and 2 researchers.

The independent variable was the type of game (an optical science education app called “Light and Color” or the AR game we designed; see Figure S2 in Multimedia Appendix 2 for a comparison of the differences between the 2 games). The dependent variables for both intervention and control group participants were the differences between the pre- and posttest results of the POE tests and the children’s motivation to play the game. To create control variables for the experiment, both games included the topics of animal vision, light transmission, and color-light mixing, and neither game involved a human teacher. In addition, there were no significant sex (P=.49) or age (P=.67) differences between the 2 groups.

**Intervention Group**

Before the test started, the researcher provided the basic information of the experiment to the participants, including the test topic, test technique, test time, and other information. The participants were asked to complete a cognitive exam on the notion of light and perform a POE test for each topic to find out how well they comprehend the content, without being told whether their answers were correct.

After completing the pretest, intervention group participants were instructed to complete the 3 game-based learning modules of the AR science education app on the iPad regarding animal vision, light transmission, and color-light mixing. On their initial encounter with the game, respondents were given around 10 minutes to comprehend its mechanics. The intervention group’s total learning time was limited to 20 minutes, the testing process was completed under the supervision of the instructor and the experimenter, and the children’s behavioral characteristics were recorded. During the experiment, the participants were not disturbed in any way; researchers only intervened when they faced difficulties or requested assistance. The participants were given a 15-minute respite at the conclusion of the trial to take another POE test. Before and after the experiment, each participant’s performance on the game was recorded. The researcher then read aloud and described the items on the intrinsic motivation and cognitive load scales to the participants, who scored the scale items using a 5-point “smiley face” scale.

**Control Group**

The control group was also introduced to the experiment and given a preintervention POE test to assess their prior knowledge of the learning material. The control group completed the same 3 game courses for a maximum of 20 minutes using the non-AR app “Light and Color” after completing the pretest. The participants took a 15-minute break at the conclusion of the trial to complete another POE test and the Intrinsic Motivation Inventory (IMI) scale (Figure 4).
**Evaluation Metrics**

The study was validated based on several experiments.

The POE test is commonly used in science classes and tries to expose students’ expectations about certain events and the rationale for these predictions [29]. It is used to demonstrate scientific experiments to pupils and is advantageous for fostering children’s critical thinking and assessing students’ grasp of scientific topics. The investigator then displays the relevant physical events to the students using basic prop materials after requiring the students to independently determine the correct answers to the questionnaire along with their justifications.

Finally, students are instructed to alter or supplement their explanations in light of the observations. Since children may appear to be able to answer the question properly but not comprehend the reasoning behind it, for each topic, it is possible that they do not comprehend the underlying concept. In this study, individuals’ accurate answers and explanations were recorded, and different situations were rated differently based on a 2-tier test [30] (Table 1). This scoring method is frequently used to evaluate students’ conceptual understanding [31]. The outcomes were categorized as correct answer+correct explanation, correct answer+incorrect explanation, incorrect answer+correct explanation, and incorrect answer+incorrect explanation. Each topic’s overall score was included in the
subsequent analysis. To avoid disruptions caused by children’s memorization of answers, the experimental posttest questionnaire in this study was different from the pretest questionnaire but was founded on the same scientific concepts. The examination topics are provided in Table S1 in Multimedia Appendix 2.

<table>
<thead>
<tr>
<th>Level of conceptual understanding</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct answer+correct explanation</td>
<td>2</td>
</tr>
<tr>
<td>Correct answer+incorrect explanation</td>
<td>1</td>
</tr>
<tr>
<td>Incorrect answer+correct explanation</td>
<td>1</td>
</tr>
<tr>
<td>Incorrect answer+incorrect explanation</td>
<td>0</td>
</tr>
</tbody>
</table>

Due to the young age of the study participants, the simplified version of the IMI adapted by Vos et al [32] was selected for this research. It was developed under game conditions with 3 subscales: interest-enjoyment, perceived competence, and effort-importance, to assess the perceived levels of motivation, enjoyment, and perceived difficulty of the participants. To investigate the negative emotions of children using the AR game, the study inserted questions from the original scale’s tension-stress section [33] (Table S2 in Multimedia Appendix 2). Participants were asked to rate the extent to which they concurred with the statement using a 5-point Likert scale depicting 5 smiling faces. A score of 5 indicated that the child participant strongly agreed with the statement.

**Ethical Considerations**

The study was approved by the Human Research Ethics Committee of Shanghai Jiao Tong University (H20220411) in China. Informed consent was signed by guardians and the data were deidentified. A toy with a value of CNY ¥50 (US $7.01) was provided as compensation.

**Results**

A total of 36 healthy Chinese children aged 6-8 years were recruited in May 2022, including 22 male and 14 female children, all of whom participated in the experiment with the consent of their guardians and of their own volition. The 36 participants were randomly assigned to the intervention group (n=18, 50%) and the control group (n=18, 50%), with the mean age of the intervention group being 7.16 (SD 0.76) years and that of the control group being 7.06 (SD 0.78) years. Baseline demographic data and POE test scores for the intervention and control groups are shown in Table 2. The statistical analysis revealed that there was no statistically significant distinction observed between the 2 groups across all variables (all P > .05). This suggests that the intervention and control groups exhibited a similar overall comprehension level prior to the commencement of the trial. The experimental procedure is provided in Figure 5.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intervention group(^a) (n=18)</th>
<th>Control group(^b) (n=18)</th>
<th>z score or chi-square (df)</th>
<th>P value(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male sex, n (%)</td>
<td>10 (56)</td>
<td>12 (67)</td>
<td>0.467 (1)(^d)</td>
<td>.49</td>
</tr>
<tr>
<td>Age (y), mean (SD)</td>
<td>7.17 (0.76)</td>
<td>7.06 (0.78)</td>
<td>−0.421(^e)</td>
<td>.67</td>
</tr>
<tr>
<td>POE(^f) test score for animal vision, mean (SD)</td>
<td>1.83 (1.10)</td>
<td>1.94 (1.21)</td>
<td>−0.296(^e)</td>
<td>.77</td>
</tr>
<tr>
<td>POE test score for light transmission, mean (SD)</td>
<td>2.83 (1.72)</td>
<td>2.67 (2.20)</td>
<td>−0.437(^e)</td>
<td>.66</td>
</tr>
<tr>
<td>POE test score for color-light mixing, mean (SD)</td>
<td>1.50 (1.09)</td>
<td>1.94 (1.16)</td>
<td>−1.031(^e)</td>
<td>.30</td>
</tr>
<tr>
<td>Total POE test score, mean (SD)</td>
<td>6.17 (2.28)</td>
<td>6.56 (2.12)</td>
<td>−0.273(^e)</td>
<td>.78</td>
</tr>
</tbody>
</table>

\(^a\)Augmented reality game.  
\(^b\)Non–augmented reality game.  
\(^c\)Mann-Whitney U test and \(\chi^2\).  
\(^d\)Chi-square value.  
\(^e\)z score.  
\(^f\)POE: predict observe explain.
The results of the normality test revealed a nonnormal distribution of the data (Table S3 in Multimedia Appendix 2). Consequently, the researchers conducted a paired-sample Wilcoxon rank sum test to compare the pre- and posttest findings of the intervention and control groups to assess any differences between the 2 groups. The results shown in Table 3 demonstrate notable fluctuations in both light transmission \((z=-2.696; P=.008)\) and total POE test scores \((z=-2.458; P=.01)\). Nevertheless, the results of the study indicate that there was no statistically significant advantage observed in animal vision \((z=-0.847; P=.42)\) and color-light mixing POE test scores \((z=-0.782; P=.46)\) as a result of the AR game intervention. It should be noted, however, that there was an improvement in scores following the intervention.
### Table 1. Between-group differences between the intervention and control groups on each of predict-observe-explain (POE) test (pre- and posttests).

<table>
<thead>
<tr>
<th>POE test score</th>
<th>Intervention group&lt;sup&gt;a&lt;/sup&gt; (n=18), mean (SD)</th>
<th>Control group&lt;sup&gt;b&lt;/sup&gt; (n=18), mean (SD)</th>
<th>Difference, mean (95% CI)</th>
<th>z score</th>
<th>P value&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Animal vision</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>1.83 (1.10)</td>
<td>1.94 (1.21)</td>
<td>0.36 (−0.71 to 1.43)</td>
<td>−.847</td>
<td>.42</td>
</tr>
<tr>
<td>Posttest</td>
<td>2.33 (1.14)</td>
<td>2.17 (0.99)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Light transmission</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>2.83 (1.72)</td>
<td>2.67 (2.2)</td>
<td>0.97 (−0.37 to 2.31)</td>
<td>−2.696</td>
<td>.008</td>
</tr>
<tr>
<td>Posttest</td>
<td>4.44 (1.76)</td>
<td>3.00 (1.88)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Color-light mixing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>1.50 (1.09)</td>
<td>1.94 (1.16)</td>
<td>0.72 (−0.44 to 1.88)</td>
<td>−0.782</td>
<td>.46</td>
</tr>
<tr>
<td>Posttest</td>
<td>2.39 (1.24)</td>
<td>2.50 (1.04)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>6.17 (2.28)</td>
<td>6.56 (2.12)</td>
<td>2.06 (−0.1 to 4.22)</td>
<td>−2.458</td>
<td>.01</td>
</tr>
<tr>
<td>Posttest</td>
<td>9.17 (2.48)</td>
<td>7.67 (1.71)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Augmented reality game.
<sup>b</sup>Non–augmented reality game.
<sup>c</sup>Mann-Whitney U test.

In this study, subjective IMI scale values acquired during the trial were statistically analyzed. It was observed that the different groups showed significant variability in levels of interest-enjoyment ($z=-2.440$; $P=.02$) and perceived competence ($z=-2.170$; $P=.03$; Table 4), whereas significant differences were not observed in levels of effort-importance ($z=-1.310$; $P=.20$) and tension-pressure ($z=-0.733$; $P=.48$).

### Table 2. Comparison of Intrinsic Motivation Inventory (IMI) variability between the intervention and control groups.

<table>
<thead>
<tr>
<th>IMI subscale</th>
<th>Intervention group&lt;sup&gt;a&lt;/sup&gt;(n=18), mean (SD)</th>
<th>Control group&lt;sup&gt;b&lt;/sup&gt;(n=18), mean (SD)</th>
<th>z score</th>
<th>P value&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest-enjoyment</td>
<td>20.28 (2.72)</td>
<td>18.00 (2.57)</td>
<td>−2.440</td>
<td>.02</td>
</tr>
<tr>
<td>Perceived competence</td>
<td>18.83 (3.05)</td>
<td>16.33 (3.34)</td>
<td>−2.170</td>
<td>.03</td>
</tr>
<tr>
<td>Effort-importance</td>
<td>12.89 (1.97)</td>
<td>11.83 (2.57)</td>
<td>−1.310</td>
<td>.20</td>
</tr>
<tr>
<td>Tension-pressure</td>
<td>10.44 (2.33)</td>
<td>9.77 (2.90)</td>
<td>−0.733</td>
<td>.48</td>
</tr>
</tbody>
</table>

<sup>a</sup>Augmented reality game.
<sup>b</sup>Non–augmented reality game.
<sup>c</sup>Mann-Whitney U test.

### Discussion

#### Principal Findings

The integration of science education into the foundational education of children aims to systematically cultivate their abilities in inductive and deductive thinking [34]. Serious games have demonstrated efficacy in enhancing teaching and learning outcomes within the contemporary domain of children’s science education [7]. AR technology has garnered growing interest in the realm of serious game design in recent times due to its ability to visually represent scientific processes that are not easily observable in real-life situations [35]. Further, the incorporation of AR technology into mobile devices has resulted in widespread adoption, facilitating the implementation of many apps [17]. Nevertheless, there is a lack of comprehensive study and experimentation to substantiate the efficacy of AR design in the realm of children’s science education. Consequently, a series of AR science instructional games were developed, focusing on the comprehension of light principles. The objective was to assess the efficacy of the games and the degree of intrinsic motivation of the students. The results showed that children who participated in the AR science game had substantially higher POE test scores and conceptual understanding of light propagation than the control group.

The study revealed that children exhibited varying levels of comprehension in relation to light concepts across diverse themes. Reliable between-group differences were detected among the topics of light propagation. The rationale behind the use of AR lies in its inherent benefits, which include the ability for children to engage in physical activity while delving into a more comprehensive exploration as compared to 2D games. Additionally, AR technology facilitates the rendering of real-world light phenomena, as supported by previous studies [36,37]. Our game was developed with the purpose of creating...
a metaphorical representation of concepts connected to light orientation using gesture-touch interactions. It aims to alleviate the cognitive burden experienced by children when trying to connect abstract information about light with real-world light occurrences. This is achieved by incorporating 3D properties of lighting effects into the game. Furthermore, a notable increase in the game scores of the intervention group was noticed across all topics. This observation serves as evidence for the beneficial influence of AR games on children’s conceptual transformation during the process of acquiring scientific knowledge. Our game also serves as a means of scientific investigation, necessitating active engagement from the children. Certain children had not before contemplated the underlying mechanisms responsible for commonplace visual occurrences. The stimulation of their drive to study and their high curiosity played a significant role in facilitating their conceptual shift and fostering the development of scientific thinking [38]. Nevertheless, the intervention group did not exhibit any noteworthy disparity compared to the control group in relation to the topics of animal vision and color-light mixing. The limited influence of the different interactive designs, specifically mixed camera mode and projected lighting mode, on children’s cognition may account for this disparity when compared to the visual representation format in the 2D game.

Intrinsic motivation is a potent factor that influences performance, learning persistence, and productivity [39]. Children in the intervention group demonstrated greater interest and enjoyment in intrinsic motivation than those in the control group, and they demonstrated an ability to embrace and comprehend the causes of certain light phenomena.

AR imparts scientific information that challenges children’s prior knowledge and stimulates their interest. Consequently, it can arouse interest in the principles and stimulate active thought [35]. During the test, we observed that participants had a keen interest in the game and avidly explored the interface’s interactive elements. Moreover, there was a significant difference in perceptual ability between the intervention and control groups. We believe that this difference stems from the fact that AR games, created by adding 3D virtual objects to real-world images, can better facilitate children’s understanding of complex concepts [15]. However, we also found that the intervention group showed some stress toward AR games. Children have a period of adjustment for things to which they are not accustomed to, as evidenced by their inattention and attempts to communicate with the observer when they encounter difficulties in the game [17]. Future research can therefore concentrate on how to provide prompts even when encountering difficulties.

The researchers observed that the children’s engagement in gameplay facilitated their conscious observation of light occurrences in their daily lives, resulting in a modest improvement in their comprehension during the final phase of the tests. Furthermore, when the optical principles pertaining to linear propagation, reflection, and refraction became increasingly complex, it became more challenging for the children to comprehend, leading to confusion in certain preintervention participants regarding the distinctions between these concepts. It is important to acknowledge that when a child misinterprets the dynamic effects, animation, or creative expression of a game feature, the game can potentially facilitate the development of novel alternative understanding. Fortunately, the occurrence of this scenario was limited in the 2 assessment tests conducted during the formal experiment.

The strengths and weaknesses of our study in comparison with other studies is shown in Table S4 in Multimedia Appendix 2.

In summary, the integration of AR into educational games has the potential to enhance children’s science education by offering a more immersive and engaging learning experience. This approach also may address the challenges associated with inadequate education and the lack of motivation among children to explore scientific subjects.

Conclusions and Limitations

The results suggest that the use of AR serious games can effectively motivate children to undergo conceptual shifts during the initial phases of science education. This, in turn, leads to an improved level of comprehension of scientific material. Furthermore, it is expected that these positive outcomes can be replicated in future preschool science education settings. This randomized controlled trial provides confirmation that the science education game we developed, using AR technology, has the potential to rectify children’s misconceptions regarding scientific concepts and improve the overall efficacy of science teaching.

However, there are also some limitations. First, the sample size used in the study was limited, and the sample population was mainly from the more resource-rich region of Jiangsu Province, China. Consequently, it is challenging to ascertain the presence of regional variations in other geographical areas. Prospective studies with large samples are needed to further confirm the results, and the results can be improved by considering gender, family upbringing, and children’s interest preferences in subsequent studies. Second, AR apps require a lot of attention and can be a distraction. It can cause students to ignore instructions or important stages of the experience. In addition, as the situation appeared in the pre-experiment, the game as a teaching tool may generate new misconceptions if the child misinterprets the content of the game. Finally, the existing game conveys scientific concepts mostly through voice-over prompts, which are insufficient to grab the children’s attention, and children may be distracted and lose essential information during the voice-over prompts.

Editorial Notice

This randomized study was only retrospectively registered, as the authors had not considered it necessary to register prospectively. The editor granted an exception from ICMJE rules mandating prospective registration of randomized trials, because the risk of bias appears to be low. However, readers are advised to carefully assess the validity of any potential explicit or implicit claims.
related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Conflicts of Interest
None declared.

Multimedia Appendix 1
The designed augmented reality game.

[MP4 File, 18278 KB - games_v12i1e47807_app1.mp4]

Multimedia Appendix 2
Supplementary tables and figures.

[DOCX File, 7270 KB - games_v12i1e47807_app2.docx]

Checklist 1
CONSORT eHEALTH Checklist.

[PDF File, 1214 KB - games_v12i1e47807_app3.pdf]

References
29. Liew CW, Treagust DF. The effectiveness of predict-observe-explain tasks in diagnosing students’ understanding of science and in identifying their levels of achievement. ERIC. 1998 Apr. URL: https://eric.ed.gov/?id=ED420715 [accessed 2023-12-07]

Abbreviations

AR: augmented reality
IMI: Intrinsic Motivation Inventory
POE: predict observe explain
Effects of a Virtual Reality Cycling Platform on Lower Limb Rehabilitation in Patients With Ataxia and Hemiparesis: Pilot Randomized Controlled Trial

Ana Rojo1,2,*, PhD; Arantxa Castrillo Calvillo3,*, BPT; Cristina López3,*, BPT; Rafael Raya1,*, PhD; Juan C Moreno2,*, PhD

*all authors contributed equally

Corresponding Author:
Ana Rojo, PhD

Abstract

Background: New interventions based on motor learning principles and neural plasticity have been tested among patients with ataxia and hemiparesis. Therapies of pedaling exercises have also shown their potential to induce improvements in muscle activity, strength, and balance. Virtual reality (VR) has been demonstrated as an effective tool for improving the adherence to physical therapy, but it is still undetermined if it promotes greater improvements than conventional therapy.

Objective: Our objective was to compare the effect on lower limb range of motion (ROM) when using VR technology for cycling exercise versus not using VR technology.

Methods: A randomized controlled trial with 20 patients with ataxia and hemiparesis was carried out. The participants were divided into 2 groups: the experimental group (n=10, 50%) performed pedaling exercises using the VR system and the control group (n=10, 50%) performed pedaling exercises without using VR. Measurements of the active and passive ROM of the hip and knee joint were taken before and after a cycling intervention, which consisted of 3 sessions of the same duration but with progressively increasing speeds (4, 5, and 6 km/h). Repeated measures ANOVAs were conducted to compare the preintervention (T_i) and postintervention (T_e) assessments within each group. Additionally, the improvement effect of using the VR system was analyzed by comparing the variation coefficient (Δ = 1 − [T_e / T_i]) between the preintervention and postintervention assessments for each group. Group comparisons were made using independent 1-tailed t tests.

Results: Significant improvements were shown in active left hip flexion (P=.03) over time, but there was no group-time interaction effect (P=.67). Passive left hip flexion (P=.93) did not show significant improvements, and similar results were observed for active and passive right hip flexion (P=.39 and P=.83, respectively). Neither assessments of knee flexion (active left: P=.06; passive left: P=.76; active right: P=.34; passive right: P=.06) nor knee extension showed significant changes (active left: P=.66; passive left: P=.92; active right: P=.12; passive right: P=.38). However, passive right knee extension (P=.04) showed a significant improvement over time. Overall, although active and passive ROM of the knee and hip joints showed a general improvement, no statistically significant differences were found between the groups.

Conclusions: In this study, participants who underwent the cycling intervention using the VR system showed similar improvement in lower limb ROM to the participants who underwent conventional training. Ultimately, the VR system can be used to engage participants in physical activity.

Trial Registration: ClinicalTrials.gov NCT05162040; https://www.clinicaltrials.gov/study/NCT05162040

(JMIR Serious Games 2024;12:e39286) doi:10.2196/39286

KEYWORDS
ataxia; cycling; hemiparesis; lower limb; neuropathology; rehabilitation; virtual reality; limb; intervention; neural; neural plasticity; therapy; muscle; strength; balance; tool; exercise; physical activity; neuroplasticity
Introduction

Background
Ataxia is an umbrella term for describing deficits in limb movement coordination such as dysmetria, dyspraxia, and dyssynergia [1]. The persistence of these deficits affects an individual’s functional ability and poses a health challenge for both patients and clinicians.

Current scientific evidence indicates that the most effective treatment for ataxia should combine balance and coordination retraining and constraint-induced functional movement therapy [2]. However, the scientific literature still lacks a consensus on the details of these interventions and the timing of their implementation to enhance the recovery of the functionality of motor deficits in an individual [3].

On the other hand, in the field of neurophysiology, it is well known that to induce changes in neuroplasticity to achieve the functional recovery of motor deficits, the application of therapies based on the repetition of movements is required [4]. Some studies point out that the principles of motor learning are directly related to the regeneration of structures and the reorganization of neuronal function [5,6]. Moreover, the amount of practice is a key factor in motor learning, as well as the feedback provided during practice [7]. In fact, physical therapists must consider both the error feedback and activity guidance as 2 fundamental components of patient interaction during therapy to promote neuromotor learning [8]. Thus, interventions that promote normal function rather than the compensation of deficits are more recommended and should be applied to generate a physical activity plan based on the principles of motor learning and neural plasticity for patients with ataxic hemiparesis.

Prior Work
The scientific literature in the field of neurorehabilitation shows that pedaling exercises have the potential to induce improvements in muscle activity, strength, and balance [9]. This is mainly due to the fact that pedaling exercises based on the use of a cycloergometer provide a high number of flexion and extension repetitions [10] in the lower extremities for considerable periods of time. Because pedaling and walking are cyclical locomotor tasks that require the lower limb to alternate between flexion and extension [11,12], both share similar locomotor patterns of alternating muscle activation of antagonists [10,13]. Thus, cycling exercises are found to be useful for strengthening the lower limb muscles while acting as a pseudowalking task-oriented exercise. Some studies eluded that those biomechanical functions may be altered by the muscle groups involved in the pedaling tasks [14-16]. In fact, it was found that the degradation of pedaling performance in adults with hemiparesis was related to abnormalities in the execution of specific biomechanical functions [15]. Subsequently, it has been proven that human walking and cycling shared similar muscle synergies [16]. This evidence is the basis for rehabilitation treatments based on pedaling movements with potential positive outcomes for walking [16].

The ergometer is an equipment designed to perform cardiovascular work based on the alternative circular movement of the lower limb. Its use is advantageous for a muscle coordination study because balance is not an applicable factor in this kinematically constrained task [13]. In fact, applying an ergometer-based cycling routine could be useful because it requires no balance. Moreover, the exercise intensity of the ergometer-based cycling can be adapted to the user by adjusting the resistance of the pedal or the target speed. The ability to personalize the intensity of the exercise is a relevant factor for the patient’s rehabilitation process. For these reasons, regular ergometer-based cycling is found to be a safer unsupervised exercise that is recommended for lower limb rehabilitation. Nevertheless, cycling exercise is also a static and repetitive form of exercise that leads to boredom and listlessness in patients. To deal with this discouragement factor, emerging technologies have been applied to elicit intrinsic motivation for rehabilitation patients [17]. Several studies pointed out the usefulness of gaming elements and virtual environments as assistive technology [18,19] and their potential effectiveness in physical therapies as opposed to conventional therapies [20].

Quite a few studies have focused on the analysis of functional metrics in virtual pedaling. A recent study evaluated the functionality of a virtual reality (VR) cycling training program that was applied to 10 patients with stroke [21]. It assessed the improvement of the bilateral asymmetry between the experimental group and the control group after the VR cycling intervention program. To evaluate this index, they equipped the ergometer pedals with force plates to determine the effect of the VR cycling training on each limb. The improvement of bilateral strength and standing balance was significantly different between VR cycling training and traditional physical training. Similarly, a previous study compared the effects of a cycling training program with extrinsic biofeedback and a nonimmersive interface versus traditional physical training on lower limb functional recovery in patients with stroke [22]. The results showed that improvements in walking endurance, walking speed, and muscle spasticity of the group using VR were significantly better than the group who underwent traditional physical training.

Objectives
The main objective of this study was to evaluate 2 different interventions: pedaling with VR and pedaling without VR. This study focused on comparing the improvements in lower limb range of motion (ROM) in pedaling activity between the group using VR and the group not using VR. To this end, a randomized controlled trial was carried out with patients with ataxia and hemiparesis. Hip and knee ROMs were measured before and after the cycling intervention. The overall aim of these analyses was to determine the effects of the 2 different interventions on short-term improvement of lower limb function and ROM.

Methods

VR System
The VR system implements extrinsic feedback strategies, gamification by levels, and personalization of the sessions with the aim of achieving greater adherence to pedaling exercise sessions. Its immersive nature means an increase in the sense of “presence,” promoting the active involvement of the user.
The VR system is based on the transmission of the cycling kinematic data captured by the inertial sensors to the Oculus Quest 2 (Meta) head-mounted display (HMD) via Bluetooth. Therefore, the virtual application estimates the pedaling cycles, cadence, and distance during the exercise activity. The VR scenarios generated for this therapy consist of mapping the cycling cadence to the vehicle speed. Thus, the patient is placed inside a vehicle and visualizes the session data on the control panel while moving at the speed of the pedaling motion.

The design of the VR experience has been technically validated computationally to ensure low latency in motion analysis and visual representation of motion [23], thus preserving the embodiment effect and the sense of presence. Subsequently, the platform has also been validated from the point of view of satisfaction and ease of use of the system [24]. Additionally, considering that it is a stationary experience with an HMD that simulates a displacement, we evaluated to which extent the VR experience generates the type of motion sickness that causes fatigue, nausea, disorientation, postural instability, or visual fatigue [25]. Indeed, we verified that the platform does not generate adverse effects due to cybersickness [24].

**Recruitment**

The participants were patients of both sexes between 18 and 90 years of age, recruited at the Lescer Clinic applying the inclusion and exclusion criteria. Inclusion criteria were as follows: individuals were eligible if they (1) had been prescribed pedaling exercise as treatment for lower limb rehabilitation and (2) were able to perform a pedaling session with VR technology. Exclusion criteria were as follows: (1) an insufficient cognitive state, (2) an unbound bone fracture, (3) severe disorders of vision or audition (inability to perceive visual or auditory information coming from VR), and (4) any incompatibility with the use of a VR system according to the clinical record. A sample of 22 participants (n=13, 59% male and n=7, 32% female; mean age 59.90, SD 13.56 y) volunteered to participate in this pilot randomized controlled trial (Table 1). Of this 22-person cohort, 1 participant dropped out of the study and 1 participant did not complete the study (Figure 1). The cohort was randomly divided into the experimental group (EG; 9/10, 90% male and 1/10, 10% female; mean age 60.80, SD 12.26 y) with VR cycling exercises or the control group (CG; 4/10, 40% male and 6/10, 60% female; mean age 59.00, SD 14.69 y) with traditional cycling exercises.

<table>
<thead>
<tr>
<th>Group and participant number</th>
<th>Sex</th>
<th>Age (y)</th>
<th>Etiology</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Male</td>
<td>57</td>
<td>Ischemic stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>71</td>
<td>Hemorrhagic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>53</td>
<td>Hemorrhagic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>72</td>
<td>MCAa stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>53</td>
<td>MCA stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>62</td>
<td>Ischemic stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>59</td>
<td>Hemorrhagic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>56</td>
<td>Progressive multifocal leukoencephalopathy</td>
<td>Ataxia</td>
</tr>
<tr>
<td>9</td>
<td>Female</td>
<td>86</td>
<td>Hemorrhagic stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>39</td>
<td>Ischemic stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td><strong>CG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Male</td>
<td>45</td>
<td>MCA stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>64</td>
<td>Hemorrhagic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>58</td>
<td>Guillain-Barré syndrome</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>4</td>
<td>Female</td>
<td>41</td>
<td>Hemorrhagic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>5</td>
<td>Female</td>
<td>49</td>
<td>Ischemic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>83</td>
<td>Ischemic stroke</td>
<td>Ataxia</td>
</tr>
<tr>
<td>7</td>
<td>Female</td>
<td>80</td>
<td>Hemorrhagic stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>8</td>
<td>Female</td>
<td>72</td>
<td>Traumatic brain injury</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>57</td>
<td>Ischemic stroke</td>
<td>Hemiparesis</td>
</tr>
<tr>
<td>10</td>
<td>Female</td>
<td>41</td>
<td>Guillain-Barré syndrome</td>
<td>Ataxia</td>
</tr>
</tbody>
</table>

*aMCA: middle cerebral artery.*

https://games.jmir.org/2024/1/e39286
Ethical Considerations

Ethical approval was obtained from the Research Ethics Committee of the San Pablo CEU University (550/21/51). This study has been registered at ClinicalTrials.gov (NCT05162040). All the participants were given written information in accordance with the Research Ethics Committee. The informed consent and the ability for participants to opt out was provided. Additionally, participants were informed that the data collected in this study can only be used for this study, not for secondary studies. The approval of the Research Ethics Committee of San Pablo CEU University only covers this study and does not cover a secondary analysis without additional consent. However, no additional analysis had been carried out.

To ensure privacy and confidentiality, data are collected by employees of the agencies participating in the study. Each participant is assigned a unique code along with personal sociodemographic data and informed consent. These files remain in the custody of the principal investigator in charge of the project, while the assigned number is the one that identifies the anonymized data that was later analyzed. Finally, the participation in this study is completely voluntary; no compensation of any nature is offered to the human participants.

Intervention

This study was designed as a randomized controlled trial with 20 participants divided into 2 groups, following a block randomization method. The participants of the EG (n=10)
performed pedaling exercises while using the VR system, whereas the participants of the CG (n=10) performed pedaling exercises without using the VR system. Before and after completing the exercise program, measurements of gait function metrics and joint ranges were performed to assess the effect of using VR stimulus during the cycling exercises.

The participants completed the cycling intervention simultaneously with their rehabilitation sessions. Afterward, for each participant, 3 cycling sessions were scheduled over 1 week with a maximum of 48 hours between sessions. Each session consisted of 2 sets of a 5-minute pedaling exercise spaced with a 2-minute break (to rest). Similar studies [19,26] have tested robotic unicycles in pedaling sessions at a cadence of 60 revolutions per minute. In our case, the pedaling speed of 1 cycle per second is equivalent to a target speed of 6 km/h. For this reason, it was decided to set this speed as the maximum speed and to start the first session with a slightly more comfortable speed (4 km/h) and increase it progressively (Figure 2). The participants of both groups performed the exercise following a set pedaling speed so that they received visual feedback according to the set target speed of 4-6 km/h for each session. The EG participants received visual feedback through the immersive VR application, whereas the CG participants received visual feedback on the ergometer display. All participants were instructed to maintain a constant pedaling speed throughout the session at the target cadence.

Figure 2. Summary of the intervention program for experimental and control group participants. VR: virtual reality.
Physical Assessment

For the assessment of active and passive ROM of the hip and knee joint, a specific ROM assessment tool was used. Measurements were extracted from biomechanical analysis using an inertial motion capture system (Werium; Werium Solutions) consisting of 2 inertial sensors: 1 placed in the distal part of the extremity (moving sensor) and the other in the proximal part (fixed sensor). Both sensors send their measurements via Bluetooth to a PC that runs the data acquisition software, Pro Motion Capture (Werium Solutions). This software computes the relative angle from both angle measurements (avoiding compensations) with an accuracy of 1 degree.

Protocol

The cycling sessions for both groups consisted of the use of a leg ergometer that allows training of the lower limb. Additionally, the EG used an inertial sensor placed on the right thigh and the Oculus Quest 2 HMD (Figure 3).

Figure 3. Cycling session of a participant in the experimental group using the virtual cycling platform.

The EG underwent the following procedure each session:

- The clinician connected the inertial sensor to the Oculus Quest 2 HMD.
- The patient was seated in a nonmovable chair (with no armrests) in front of the pedaling station during the entire session. The inertial sensor was placed on the right thigh.
of the patient by adjusting an elastic band, and the sensor was turned on.

- The clinician fitted the Oculus Quest 2 HMD comfortably on the patient and guided him or her through the selection of the game scene. Once the game environment was entered, the clinician indicated the number of minutes of exercise and the target speed of the session so that the patient could configure these parameters on the interactive settings panel.

- Finally, the user performed 2 sets of a 5-minute cycling exercise with a 2-minute break between the sets.

Similarly, the CG underwent the following procedure each session:

- The patient was seated in a nonmovable chair (with no armrests) in front of the pedaling station during the entire session.

- The clinician turned on the ergometer’s display and entered the number of minutes of exercise and the target speed of the session.

- Finally, the user performed 2 sets of a 5-minute cycling exercise with a 2-minute break between the sets.

Statistical Analysis

The data analysis model is the repeated measures model between 2 groups and the analysis of the longitudinal effect in increments of the measurements. Multifactor ANOVA analysis (with P<.05) were computed with SPSS Statistics (version 27.0; IBM Corp). The sample size was calculated using the software tool G*Power (version 3.1.9.7; Heinrich Heine Universität Düsseldorf). Ideally, assuming an effect size of 0.7, a minimum sample of 20 participants was required for the study to provide consistent statistical results. Since the effect size shows the strength of the relationships, it represents a minimum clinically meaningful difference. Of the many different types of effect sizes, the G*Power software uses Cohen $d$ to characterize effect size by relating the mean difference to variability. Therefore, his study standardized the effect size to 0.7 for sample size calculation and power analysis.

Results

To identify the underlying differences between the preintervention ($T_i$) and postintervention ($T_e$) assessments in each group, repeated measures ANOVAs were conducted with time ($T_i - T_e$) as the dependent variable and group as the main within-subjects factor. When the ANOVA was significant, the Bonferroni post hoc test was used. To ensure that the error variance of the dependent variables is equal across groups, the Levene test was applied beforehand for all the metrics.

In addition, to identify the improvement effect due to the use or nonuse of the VR system, the variation coefficient between the preintervention and postintervention assessments was analyzed for each group as follows: $\Delta = 1 - (T_e / T_i)$. The variation coefficient outcomes were compared between groups by the independent 1-tailed $t$ test. The mean and SD of the ROM outcomes for the hip and knee of each group are shown in Table 2. The mean increase $\Delta$ for each measurement is shown in Figures 4 and 5.
Table. Hip and knee range-of-motion outcomes.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Control group, mean (SD)</th>
<th>Experimental group, mean (SD)</th>
<th>Variation coefficient (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Preintervention (°)</td>
<td>Postintervention (°)</td>
<td>Preintervention (°)</td>
</tr>
<tr>
<td>ALHF(^a)</td>
<td>81.25 (36.09)</td>
<td>94.23 (32.26)</td>
<td>92.84 (21.40)</td>
</tr>
<tr>
<td>PLHF(^b)</td>
<td>106.07 (21.16)</td>
<td>107.94 (17.63)</td>
<td>112.92 (17.76)</td>
</tr>
<tr>
<td>ARHF(^c)</td>
<td>97.55 (20.94)</td>
<td>97.13 (21.26)</td>
<td>97.11 (28.05)</td>
</tr>
<tr>
<td>PRHF(^d)</td>
<td>106.63 (17.06)</td>
<td>109.82 (14.99)</td>
<td>119.74 (14.73)</td>
</tr>
<tr>
<td>ALKF(^e)</td>
<td>46.07 (14.62)</td>
<td>45.97 (11.47)</td>
<td>37.47 (12.03)</td>
</tr>
<tr>
<td>PLKF(^f)</td>
<td>58.82 (9.84)</td>
<td>55.96 (9.79)</td>
<td>57.14 (13.92)</td>
</tr>
<tr>
<td>ARKF(^g)</td>
<td>39.13 (16.54)</td>
<td>37.81 (10.68)</td>
<td>43.03 (10.00)</td>
</tr>
<tr>
<td>PRKF(^h)</td>
<td>50.57 (10.02)</td>
<td>49.81 (10.31)</td>
<td>63.35 (12.28)</td>
</tr>
<tr>
<td>ALKE(^i)</td>
<td>61.72 (14.86)</td>
<td>62.92 (13.11)</td>
<td>55.57 (17.13)</td>
</tr>
<tr>
<td>PLKE(^j)</td>
<td>66.46 (11.74)</td>
<td>69.95 (15.09)</td>
<td>64.75 (11.94)</td>
</tr>
<tr>
<td>ARKE(^k)</td>
<td>64.00 (10.11)</td>
<td>68.02 (10.14)</td>
<td>57.49 (14.91)</td>
</tr>
<tr>
<td>PRKE(^l)</td>
<td>66.67 (11.53)</td>
<td>67.18 (10.93)</td>
<td>57.65 (11.21)</td>
</tr>
</tbody>
</table>

\(^a\)ALHF: active left hip flexion.  
\(^b\)PLHF: passive left hip flexion.  
\(^c\)ARHF: active right hip flexion.  
\(^d\)PRHF: passive right hip flexion.  
\(^e\)ALKF: active left knee flexion.  
\(^f\)PLKF: passive left knee flexion.  
\(^g\)ARKF: active right knee flexion.  
\(^h\)PRKF: passive right knee flexion.  
\(^i\)ALKE: active left knee extension.  
\(^j\)PLKE: passive left knee extension.  
\(^k\)ARKE: active right knee extension.  
\(^l\)PRKE: passive right knee extension.
Figure 4. Summary of increments in active and passive hip ROM parameters with SD bars. The vertical axis represents the percentage of postintervention increase or decrease of each hip ROM parameter. ALHF: active left hip flexion; ARHF: active right hip flexion; PLHF: passive left hip flexion; PRHF: passive right hip flexion; ROM: range of motion.
Figure 5. Summary of increments in active and passive knee ROM parameters with SD bars. The vertical axis represents the percentage of postintervention increase or decrease of each knee ROM parameter. ALKE: active left knee extension; ALKF: active left knee flexion; ARKE: active right knee extension; ARKF: active right knee flexion; PLKE: passive left knee extension; PLKF: passive left knee flexion; PRKE: passive right knee extension; PRKF: passive right knee flexion; ROM: range of motion.

With regard to the hip flexion outcomes, the active left hip flexion results were significant by ANOVA (P=.03), with no significance observed for the between-subjects effects test (P=.67). However, the within-subjects effects test was significant for the time factor (P=.03), but no significant group-time interaction effect was found (P=.08). Despite the opposing results showing passive left hip flexion improvements for each group, there was no significance difference by ANOVA (P=.93) and no statistically significant result was obtained by the between-subjects effects test. Passive left hip flexion was statistically significant in the within-subjects effects test for the time factor (P=.008). The active and passive right hip flexion results were not significant by ANOVA (P=.39 and P=.83, respectively). In both cases, no significant results were obtained for the between- and within-subjects effects tests.

For the knee ROM measurements, when analyzing the left knee assessments, the active and passive left knee flexion outcomes were not significant by ANOVA (P=.06 and P=.76, respectively). No statistically significant results were obtained by the between- and within-subjects effects tests in both cases. Similar results were obtained for the active left knee extension outcomes. Although reasonable differences in the active and passive left knee extension increases between groups can be
the significant by ANOVA (P<0.05). No statistically significant results were obtained by the between- and within-subjects effects tests in both cases.

Regarding the right knee assessments outcomes, all outcomes were not significant by ANOVA (active flexion: P=0.34; passive flexion: P=0.06; active extension: P=0.12; passive extension: P=0.38). No statistically significant results were obtained by the between- and within-subjects effects tests for all cases, except for passive right knee extension, which was statistically significant for the time factor (P<0.05) by the within-subjects effects test.

Discussion

Principal Findings

The aim of this study was to test the short-term effects of 2 different interventions on short-term improvement of lower limb function and ROM. For this purpose, a randomized controlled trial was carried out with participants with ataxia and hemiparesis.

In this study, the improvement outcomes of active and passive knee and hip joint ROMs due to the use of VR technology were inconclusive. Likewise, no statistically significant differences in the results between groups can be indicated. Even so, all the active ROMs measured—that is, performed by the patients—showed an increase with respect to the initial values. A greater disparity was observed in the passive measurements, although this may be attributed to the different passive mobilizations performed at each time by different physiotherapists. In this case, the active measurement is of special relevance in clinical terms because it indicates a ROM that the patient is able to achieve autonomously. On the other hand, large SDs in outcome variables clearly indicate that the improvements in the functional gait outcomes are not entirely consistent or represent a group effect. We observe that no significant effect can be attributed to VR intervention based on the statistical analysis of the immediate effects on gait function and joint ROM.

However, considering this similarity between groups, it can be pointed out that the use of VR has similar positive effects as the use of the conventional pedaling treatment. Thus, this immediate observation of effects leads us to conclude that the use of VR during pedaling exercise has similar effects to non-VR exercise training. Therefore, given that the use of VR technology does not worsen the improvement of lower limb ROM, and in line with the scientific literature [17-20], it may be advantageous to use it to maintain the patient’s motivation.

Strengths and Limitations

A limitation of this study is the short-term nature of the intervention program. It is arguable that a longer intervention program would have shown more notable effects on functional improvement. However, assuming that it is precisely the treatment time that is one of the main causes of progress in physical improvement, the motivational impact of VR technology over time would need to be assessed. Therefore, further studies on the motivational impact of VR cycling versus conventional cycling on long-term physical activity remain to be addressed. Regarding these future studies, we suggest that cohort studies should be conducted among a population with more homogeneous neurological conditions. This recommendation is based on the limitations encountered in this study, where the difficulty of drawing conclusions about group changes or improvements with such wide SDs is presumably a reflection of the heterogeneity of the group.

Another factor to consider is that different physiotherapists were involved in taking the ROM measurements of the participants, although the measurement system was the same. This fact could be considered in future studies to evaluate interrater effects.

Future Directions

We consider it relevant to analyze, in future studies, whether these improvements in active and passive ROM are accompanied by greater muscle activation, in particular, the hamstrings, rectus femoris, gastrocnemius, and tibialis anterior muscles, as suggested by scientific literature [27].

Conclusions

The results of this trial demonstrate that pedaling exercises coordinated with VR technology works as successfully as conventional training for patients with lower limb disorders such as ataxia and hemiparesis. In this study, it was found that participants who performed the pedaling exercise program using the VR system showed similar results to the participants who performed the exercise activity without using VR technology. Overall, VR technologies can be a useful tool to help patients with ataxia and hemiparesis engage in lower limb exercise therapies.

Acknowledgments

The authors would like to thank all the participants who collaborated in this study, as well as the therapists and health care professionals from Centro Lescer for their participation in this study. The financial support for the industrial doctorate project “Desarrollo y estudio de una plataforma interactiva y un sistema electrónico de pedaleo para rehabilitación funcional de personas mayores” of the Autonomus Community of Madrid (IND2019/TIC17090) toward this research is hereby acknowledged. Grant PID2021-127096OB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by “ERDF A way of making Europe.”

The funding sponsors have no role in the design of the study; the collection, analyses, or interpretation of data; the writing of the manuscript; and the decision to publish the result.
Data Availability
The data sets generated or analyzed during this study are available on the GitHub repository [28].

Authors' Contributions
AR contributed to software, data curation, formal analysis, and writing--original draft. ACC contributed to data curation and methodology. CL contributed to methodology, resources, and supervision. RR contributed to funding acquisition, supervision, and writing--review and editing. JCM contributed to funding acquisition, supervision, and writing--review and editing.

Conflicts of Interest
RR is the chief executive officer of Werium Solutions, and AR is a software developer at Werium Solutions. The other authors declare no conflicts of interest.

Checklist 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6.1).

References

Abbreviations
CG: control group
EG: experimental group
HMD: head-mounted display
ROM: range of motion
VR: virtual reality

© Ana Rojo, Arantxa Castrillo Calvillo, Cristina López, Rafael Raya, Juan C Moreno. Originally published in JMIR Serious Games (https://games.jmir.org), 4.1.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
Exploring the Use of a Learning-Based Exergame to Enhance Physical Literacy, Soft Skills, and Academic Learning in School-Age Children: Pilot Intervventional Study

Aurelie Goncalves, PhD; Florence Lespiau, PhD; Gaëtan Briet, PhD; Eugénie Vaillant-Coindard, MSc; Angèle Palermo, MSc; Elsa Decobert, BSc; Nathan Allegret-Bourdon, BSc; Elodie Charbonnier, PhD

APSY-v, University of Nîmes, Nîmes, France

Corresponding Author:
Aurelie Goncalves, PhD
APSY-v
University of Nîmes
Rue du docteur Georges Salan
CS13019
Nîmes, F-30021
France
Phone: 33 466364518
Email: aurelie.goncalves@unimes.fr

Abstract

Background: There is ample evidence that most children do not perform enough physical activity (PA). To address this major public health problem, the French government implemented 30 minutes of daily PA (DPA) at schools but did not provide any supplemental resources or concrete guidance. Considering both children’s interest in video games and the need for teachers to complete their curriculum, the use of a learning-based exergame that combines PA and learning appears particularly relevant.

Objective: The first objective of this study was to evaluate the feasibility of implementing 30 minutes of DPA through exergaming among school-age children. The second objective was to examine the effects of an exergaming program on physical literacy, academic learning, and soft skills (motivation, self-efficacy, and concentration).

Methods: This interventional study had a pre-post design and used the Play LÜ exergame platform. The study included 79 children with a mean age of 8.9 (SD 1.2) years from grade 2 (7 years old) to grade 5 (11 years old). Play LÜ requires players to throw balls against a wall to reach a target or to activate an object and provides an interactive game area for educational activities linked to specific learning themes. After a 4-session familiarization phase during which the teachers chose to prioritize mathematics learning in 30-minute DPA sessions, students took part in DPA sessions over a period of 3 weeks with Play LÜ and a motor skills circuit behind the LÜ setup to keep them continuously active. All sessions were carried out by PA specialists. Each session started with a warm-up using the Grööve application, continued with main activities promoting mathematics learning adapted to each grade level, and ended with a 3-minute meditation for returning to a calm and serene state using the Gaïa application. Before (T0) and after (T1) the program, students completed a self-evaluation booklet to assess their levels of physical literacy, academic performance, and soft skills.

Results: The implementation of this exergaming program was welcomed by the school’s administration, teaching staff, and parents. After the program, we observed increased scores for physical literacy (difference +2.6, percentage change +3.6%; W=933.0; P=.002; rṇ=−0.39, 95% CI −0.58 to −0.16) and motivation in mathematics (+0.7, +9.8%; W=381.5; P=.005; rṇ=−0.44, 95% CI −0.66 to −0.16). In addition, it is important to note that some measures progressed differently across learning levels and age groups.

Conclusions: The study results indicate positive impacts of learning-based exergaming on physical literacy and motivation in mathematics among school-age children.

(JMIR Serious Games 2024;12:e53072) doi:10.2196/53072

KEYWORDS
learning support; exergaming; physics playground; educational games; primary school; children
Introduction

Background

Regular exercise and physical activity (PA) have been shown to benefit children’s physical and social health, as well as their academic performance [1-3]. For children, the World Health Organization (WHO) recommends a minimum PA practice of 60 minutes of daily moderate-to-vigorous PA [4]. Yet, in France, only 41.8% of children reach these recommendations [5]. In other words, nearly 6 out of 10 children are physically inactive. In this context, the French national health strategy [6] has set as a major objective the implementation of a comprehensive policy of prevention and health promotion. As a result, since September 2022, primary schools have been required to provide 30 minutes of daily PA (DPA) to promote PA and encourage the development of children’s motor skills and physical abilities. Distinct from the teaching of physical education (PE), the 30 minutes of DPA can take a variety of forms, adapted to the context of each school. They can be split up and combined over the various school and extra-curricular periods. This was extended to all elementary schools after 2 years of testing in 11,000 volunteer schools [7].

Given the significant amount of time children spend in school throughout their childhood, schools represent an ideal setting to achieve maximum impact with regard to improving PA levels [8,9]. Furthermore, several studies have suggested that combining PA with academic activities can improve children’s health and cognitive functioning, which could subsequently lead to an improvement in children’s academic performance [10-12]. Attention, in particular, which is a prerequisite for learning, is often targeted during classroom-based PA [13]. However, other variables, such as motivation and self-efficacy (referring to the child’s perception of his or her capabilities), are well known to influence children’s school performance and acquisition [14-17] and could be influenced by a more entertaining approach to learning via exergaming.

Despite all the benefits of implementing regular PA at school, in a systematic review, Nathan et al [18] highlighted several barriers, such as environmental context and resources with “a lack of time in the curriculum;” goals with “competing curriculum demands of other subjects” or “physical activity considered a lower priority than other subjects;” and beliefs about capabilities, such as a lack of teacher expertise and confidence in delivering PA, and intentions with “a lack of teacher motivation to implement PA” [18]. By contrast, the authors also mentioned several facilitators. Among them, the knowledge domain was indicated to play a facilitating role, for example, “sufficient knowledge about PA and health to effectively conduct PA” [18]. This dimension could be explored through the notion of physical literacy (PL), which corresponds to “the motivation, confidence, physical competence, knowledge, and understanding to value and take responsibility for engagement in physical activities for life” [19]. Indeed, PL is particularly important in early childhood, a crucial period for the development of fundamental movement skills [20] and the adoption of PA habits. Physically literate individuals are more physically active, spend more time playing sports, and are less sedentary. PL is a multi-level concept that is increasingly taken into consideration in the field of public health as it is a key determinant of PA habits across the lifespan [21].

One of the factors behind children’s low levels of PA and high levels of sedentary behaviors is screen time use. Indeed, 71.7% of French boys and 58.5% of French girls aged 6 to 10 years have more than 2 hours a day of screen time [5]. Children aged between 8 and 12 years have 1.5 hours of daily screen time attributable to video games [22]. Over the last decades, video games have emerged as one of the most popular forms of global entertainment. Given children’s keen interest in video games, it seems particularly appropriate to use gamification to encourage PA. For staying active while enjoying the pleasures of video games, a worthwhile alternative is exergaming. Indeed, exergaming or active video gaming requires bodily movements to play the game and encourages PA, with a focus on children’s interest in the game’s dynamics and stimulation. Our approach to exergaming takes into account a health dimension and can be associated with the conceptualization proposed by Oh and Yang [23]; defining an exergame as “a video game that promotes (either via using or requiring) players’ physical movements (exertion) that is generally more than sedentary and includes strength, balance, and flexibility activities.”

Given that children spend most of their time at school, that they have a particular appeal for video games, and that exergaming seems to have beneficial effects on school learning [24], the use of exergaming at school appears to be an ideal solution for promoting PA, PL [25-27], and learning [28]. Furthermore, it appears that the use of a technology-based learning environment at school can increase soft skills, such as motivation and concentration on academic tasks [29]. Similarly, it has been shown that incorporating technology into an instructional intervention can improve students’ sense of self-efficacy [30], which is a key variable for academic learning. Exergames, in particular, have been found to promote cognitive functions, motor skill training, enjoyment, and motivation to play among school-age children [31], and improve self-efficacy over traditional exercises [32]. Supporting this idea, it has been shown that exergaming (eg, Nintendo Wii Games [33]) incorporated into PE classes combined with health messages has a higher potential to enhance PA-related attitudes and behaviors than regular PE classes, especially in elementary school children [27]. An interesting exergaming tool for reconciling learning and DPA is the Play LÚ exergame platform (LÚ Interactive Playground) [34]. This technology can be used to change the traditional sports-school atmosphere into an interactive learning environment through interactive wall projection and a synchronized sound system. LÚ Playground activities are designed to improve the learning of children and adolescents by allowing them to respond to questions in specific fields (eg, mathematics, history, and natural sciences) by throwing balls against an interactive wall. This tool would therefore allow the practice of PA within non-PE curricula and thus ensure the 30 minutes of DPA among primary school children. Moreover, given the associations among cognitive functioning, soft skills, learning, and PA demonstrated in the literature, it appears essential to assess whether an exergaming program can improve these different variables.

https://games.jmir.org/2024/1/e53072
Objectives, Research Questions, and Hypotheses

The first objective was to study the feasibility of implementing 30 minutes of DPA through exergaming. Given that exergaming combines the interests of children (for video games) and teachers (for learning and respecting the curriculum) while promoting PA, we hypothesized that it will enable effective implementation of the 30 minutes of DPA in schools.

The second objective of this study was to evaluate the effects of an exergaming program on PL, academic learning, and soft skills (motivation, self-efficacy, and concentration). We hypothesized that implementing a DPA program involving exergaming on a specific academic course combined with information on health-promoting behaviors daily could increase children’s PL (hypothesis 1) and increase academic performance (hypothesis 2). Indirectly, allowing students to work on an academic subject more entertainingly through exergaming could improve students’ motivation in the academic discipline (hypothesis 3), their sense of self-efficacy in the subject (hypothesis 4), and their concentration in class and the academic subject (hypothesis 5).

Methods

Population

This study was conducted with children aged 7 to 11 years as part of the implementation of the 30 minutes of DPA policy. The study was designed as an interventional study with a pre-post design. It included children from grade 2 (7 years old) to grade 5 (11 years old) in a mid-sized city school in the southern part of France, who had never benefited from any intervention in the field of exergaming. Before the project, the study and objectives of this research were presented to the school administration and then to the teachers. This pilot study took place in a small school with 1 class per level and 1 teacher per grade, with each of them (n=4) having no experience of exergaming and volunteering to take part in the research protocol. This school was selected for its pre-existing collaboration with the research team and middle-school students (8th grade), as well as for the availability of a space that could be used to install the LÜ mobile setup over a period of several weeks.

Subsequently, the parents of the children in the classes concerned were informed that their children would be part of a research protocol on 30 minutes of DPA during school time. A request for parental consent was sent via the school administration to each parent. In the event of parental refusal (only 3 parents refused), the children’s data were not analyzed. A habituation phase was then proposed, and the teachers were able to learn about the various potentialities of the LÜ tool, as well as the implementation of the 30 minutes of DPA by the project team. The intervention then began and lasted 3 weeks, and preintervention (T0) and postintervention (T1) assessments were conducted.

During the enrollment period, 102 children were eligible (Figure 1). However, owing to the absence of parental consent (n=3) or the absence of children at evaluation time 0 (n=6) or time 1 (n=11), the analyses were carried out on 79 children. This final sample was made up of 34 girls and 45 boys, with a mean age of 8.9 (SD 1.2) years.
Class Measures
All the teachers expressed the wish to work on mathematics (geometry and arithmetic). A planning schedule was drawn up with the classes concerned so that the sessions could be scheduled during mathematics lessons.

Teacher Measures
At the end of the program, teachers were asked the following questions: On a scale from 0 to 10, how would you rate (1) your students’ motivation for mathematics before the program? (2) your students’ concentration for mathematics before the program? (3) your students’ motivation for physical activity before the program? (4) your students’ motivation for mathematics today? (5) your students’ concentration for mathematics today? (6) your students’ motivation for physical activity today?

Child Measures
At T0 and T1, students completed a questionnaire consisting mainly of analog visualization scales or checkboxes on different variables of interest (PL, motivation, self-efficacy, and concentration), which are described in the following sections. In addition, exercises adapted according to grade level were proposed in the target subject (mathematics) and a control subject (French).

Physical Literacy
PL was assessed using the Physical Literacy Assessment for Youth Self (PLAYself), designed for children aged 7 years or
older, to explore children’s perceptions of their PL [35]. PLAYself demonstrated robust psychometric properties, with good fit statistics, internal reliability, and a lack of item bias and problematic local dependency [36]. For a better understanding of the different dimensions of PL in the PLAYself questionnaire, the forms are available in English [37] and French [38] versions. The adaptation of this form within the evaluation booklet of this pilot project is available in Multimedia Appendix 1.

PLAYself consists of 22 questions divided into the following four subsections: (1) **Fitness**, which involves children's perceived fitness level with “disagree” and “agree” response categories for a single item; (2) **Environment**, which involves measures of 6 different environments in which children can do sports and activities (eg, “How good are you at doing sports and activities in the gym?”) on a 5-point Likert scale ranging from 1 (“never tried”) to 5 (“excellent”); (3) **Physical literacy self-description**, which involves 12 statements about doing sports and activities based on cognitive and affective factors (eg, “It doesn’t take me long to learn new skills, sports, or activities”), where the children are asked to rank how well they agree on a 4-point Likert scale ranging from 1 (“not true at all”) to 4 (“very true”); and (4) **Relative ranking of literacies**, which involves children’s ranking of the importance of literacies in school, at home with family, and with friends (eg, “Math and numbers are very important in school”) on a 4-point Likert scale ranging from 1 (“strongly disagree”) to 4 (“strongly agree”).

The first section is informative, while in the other 3 subsections, a separate score can be calculated and a total score can be obtained for PLAYself. The total PLAYself score is the average across the scores of each subsection, excluding the fitness question. A higher score (range 0-100) indicates a higher self-perceived PL.

### Academic Achievement

To measure academic achievement, exercises in French and mathematics were retrieved from the national program by level following teacher school year progression. The test evaluated students’ academic knowledge and skills related to specific subject areas, including French and mathematics. The test was grade-specific; did not contain any bias regarding age, gender, or ethnicity; and was scored as a percentage of achievement in French on one side and mathematics on the other.

### Motivation, Self-efficacy, and Concentration

Motivation and self-efficacy were assessed by 2 items each, one for mathematics and the other for French. For motivation, children were asked: “How much do you enjoy doing [mathematics/French] exercises?” For self-efficacy, children were asked: “How well do you think you did on the [mathematics/French] exercises?” Concentration was assessed by 3 items, one for mathematics, one for French, and a more general one targeting concentration in class. For this variable, children were asked: “How easy is it to concentrate in [class/mathematics/French]?” We used a simple question per variable to reduce the time needed to complete the entire protocol. The items were formulated as clearly as possible to be adapted to the children’s age and to ensure that they measure the core component of each variable. For all items, children were asked to respond using a 10-cm–long visual analog scale representing their feelings and marked by extreme labels at 0 cm (eg, very hard) and 10 cm (eg, very easy), which appeared as reliable response options in children’s questionnaires [39].

### Procedure

#### Habituation Phase

On Thursdays in March 2023, students had 4 LÜ 30-minute habituation sessions, spaced 1 week apart, enabling them to familiarize themselves with the interactive gymnasium. Activities linked to the academic development of the LÜ catalog were proposed, targeting language (ie, Minewörd), mathematics (ie, Wilk, Newton, Constello, and SphYnX), science and technology (ie, Brùsh and Grùb), history (ie, Stòria), and arts (ie, Pixël). With mathematics accounting for one-fifth of the school program in each grade and the LÜ catalog offering more mathematics-related applications (except PE, which was not at the center of the project), the 4 teachers wanted to work on mathematics during the 3-week DPA immersion phase (Figure 2).
**Immersion Phase**

After the habituation phase, students took part in 30-minute DPA sessions using the Play LÜ exergame platform and worked on a single subject selected by the teacher (Figure 2).

**Exergame Setting**

For this research protocol, the LÜ mobile equipment owned by the research team was made available to the school for this pilot project and was installed in a designated space for the duration of the project. The Play LÜ exergame platform (LÜ Interactive Playground) has the potential to overcome the limitations of a physical room. With Play LÜ, the participants are immersed in the games displayed on a giant projection wall (6×3 m). The principal mechanism of Play LÜ requires the players to throw balls against the wall (eg, to reach a target or to activate an object; Step 1 in Figure 3). In addition, this mobile platform offers an interactive game area for educational activities linked to specific learning themes (calculations, puzzles, etc). For this research protocol, a work area with a daily changing activity circuit was implemented behind the LÜ mobile setup (without the interactive wall). With class compositions ranging from 26 to 30 students, this “with” and “without” interactive wall configuration was essential to keep students active during the 30-minute session. Once the ball was sent to the interactive wall (Step 1), the student was required to go behind the LÜ mobile setup toward the back of the room (Step 2) to carry out various exercises to promote different motor skills (eg, jumping, throwing, and balancing) and perform other exercises on the way back (Step 3). At the end of the circuit behind the LÜ setup (Steps 2 and 3), the student waited for his or her turn in front of the interactive wall (Step 4).

---

**Figure 2.** Timeline of the project.

**Evaluation Phase I**
- Self-evaluation booklet completion
- **March 2023**
- **March 1**

**Habituation phase**
- Four 30-minute sessions

**Immersion phase**
- From Monday to Friday (except Wednesday no school)
- 11 sessions
- **April 3 to 20**
- **April 21**

**Evaluation Phase II**
- Self-evaluation booklet completion
Daily Session Exercises With Play LÜ

During the 3-week immersion phase, sessions were structured in the same way, with a warm-up using the Grööve application at the beginning (Figure 4), which can be assimilated with the active video game Just Dance (Nintendo) [40], and then a core session promoting mathematics learning adapted to each grade, involving a section in front of the interactive wall with the Newton application for arithmetic and Puzz application for geometry (ie, with picture geometric forms or rules), and a section without the interactive wall consisting of a motor skills circuit (eg, throwing, jumping, and balancing) that enabled the child to be as active as possible (Figure 4). During break times (mainly while waiting for their turn on the interactive wall), the children had access to posters presenting active health behaviors with their favorite heroes according to age (ie, The Minions, Miraculous, and a successful French singer or Youtuber, depending on student age). The session ended with a 3-minute meditation for a return to peace and quiet, using the Gaia application (Figure 4).
Figure 4. Daily physical activity core session details.

1. Warm-up

2. Mathematics core session

+ Motor skills circuit in background

3. Peaceful & quiet return

Focus on LÜ Applications
The Play LÜ exergame platform allows the use of applications (with or without customization) that can be used to meet general or specific learning objectives. The applications used during the immersion phase and their pedagogical benefits are summarized in Table 1.

Table 1. Play LÜ applications.

<table>
<thead>
<tr>
<th>Application</th>
<th>Duration (min)</th>
<th>Description</th>
<th>Learning objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Academic development</td>
</tr>
<tr>
<td>Groöve</td>
<td>4</td>
<td>This is a perfect warm-up and allows the development of gross motor skills.</td>
<td>● Movement</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dance</td>
</tr>
<tr>
<td>Newton</td>
<td>10</td>
<td>Newton is a fun way to combine physical activity and mathematics with the customization of equations by grade.</td>
<td>● Arithmetic</td>
</tr>
<tr>
<td>Puzz</td>
<td>10</td>
<td>Throw the ball at a piece of the puzzle to rotate it and allow it to create an active knowledge competition.</td>
<td>● Puzzle created with geometry-related images</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Spatial orientation</td>
</tr>
<tr>
<td>Gaia</td>
<td>3-5</td>
<td>This application helps students cool down after they have been active.</td>
<td>● Health and healthy habits</td>
</tr>
</tbody>
</table>

The customization of pre-existing games in LÜ was carried out on the Newton and Puzz applications for mathematics learning. For the Newton application, the difficulty of the operations proposed depended on each grade (e.g., addition and subtraction calculations for grades 2 and 3, and multiplication calculations for grades 4 and 5). For the Puzz application, the images to be assembled were linked to the geometry program (e.g., polygons...
and nonpolygons for grades 2 and 3, and complex polygons for grades 4 and 5).

**Evaluation Phase**

Before (T0) and after (T1) the 3 weeks of DPA (Figure 2), students completed a self-evaluation booklet to assess their levels of PL, academic performance, and soft skills that could be impacted by the program (eg, motivation, concentration, and self-efficacy).

At the end of the 3-week immersion phase, the students retook the same questionnaire, with only minor changes to the exercises (eg, 12+18 replaced by 14+15), but with the same instructions and level of difficulty. Following this, a short interview was conducted with the teachers, asking them to assess the changes observed in their classrooms.

All DPA sessions were carried out by sports science students specializing in adapted PA and health, under the supervision of a qualified teacher of adapted PA and health.

**Statistical Analysis**

Power analyses were conducted using G*Power (Heinrich-Heine-Universität Düsseldorf). For a pre-post comparison, with a medium effect size (0.50), an alpha error probability of .05, and a power of 0.95, we obtained a total sample size of 45. We then adjusted according to the number of participants available in the school, which allowed us to reach the sample size of 45.

As our data did not follow a normal distribution and given the characteristics of our sample, the intervention impact was tested with nonparametric within-group comparisons (T0 vs T1; Wilcoxon test, bilateral P values) for all participants and then by school-grade grouping. Effect sizes were expressed as the rank biserial correlation ($r_{bi}$) and its 95% CI. We also provided the score differences between T1 and T0, and expressed them as a percentage of improvement. Data were analyzed using JASP software (version 0.17.2.1; JASP Team).

**Ethical Considerations**

This pilot study involved an experiment in human and social sciences in the field of health. As mentioned in article R1121-1 section II subsection D of the French Public Health Code, this type of experimentation in human and social sciences does not require the authorization of the Committee for the Protection of Persons. Before the start of the study, a favorable opinion was obtained from the president of the University of Nimes ethics committee. This individual verified that the study was conducted in accordance with institutional and national ethical standards, as well as the Declaration of Helsinki (2008). Moreover, this study was integrated into the school’s activities and projects, and the protocol was validated by the school administration.

**Consent to Participate**

Concerning consent and information, the study was first presented to the school’s teaching staff and administration who gave their approval to take part. Next, an online information notice and online informed consent form with the names and university affiliations of the experimenters were provided to the parents of all children in the classes involved in the study before initiation. Finally, the information and informed consent of the children and their teachers were collected face-to-face. Recruitment was based on voluntary participation, with no compensation for participants. Participants were informed that they could withdraw their consent at any time, whether at the request of the child, parent, or teacher.

**Specific Measures Taken**

To assure safety and security, all activities took place during class hours, under the supervision of the teacher, and the exercises were led by an associate professor specializing in public health and PA and two 3rd-year students in adapted PA from the University of Nimes. The expertise of the 3 animators enabled them to adapt the PA to the children’s abilities in order to prevent any risk of injury. Moreover, the number of animators made it possible to provide individual support when needed. Finally, to guarantee the security of the data, they were stored on a secure university computer, and the printed versions of the data were kept in a secure cupboard in a university office.

**Results**

**Feasibility of DPA and Exergaming**

With regard to our first objective, which was to study the feasibility of implementing 30 minutes of DPA through exergaming, our results showed that implementing exergaming during school time is entirely feasible. First, regarding the parents, all but 3 were in favor of their children taking part in the project. Second, regarding the teachers, all agreed to take part in the project. Third, all scheduled sessions (n=11) were carried out, with no sessions canceled. External constraints, such as educational visits or other activities, could have led to cancellations, but the teachers expressed a desire not to miss any sessions and agreed to exchange schedules with other classes when constraints arose, demonstrating their interest.

**Effects of the Exergaming Program**

**Effects on the Entire Cohort**

With regard to our second objective for the whole cohort, the 3-week intervention of DPA led to increased scores in PL (in accordance with hypothesis 1) and motivation in mathematics, which was the subject covered in the intervention (in accordance with hypothesis 3). There was a general improvement regarding concentration in class, and we expected (hypothesis 5) this increase to be observed for mathematics as well (Table 2). In addition, contrary to our assumptions, we did not observe any changes in academic performance (hypothesis 2) or feelings of self-efficacy (hypothesis 4) in mathematics. Surprisingly, the intervention also favored French learning, which was not covered in the intervention, with academic performance, concentration, and self-efficacy in French being higher after the intervention.
Effects on Grade 5 Participants

Focusing specifically on each grade (see Multimedia Appendix 2 for full details), participants in grade 5 were those most affected by the intervention.

After the intervention, grade 5 participants showed an increase in PL (difference +5.6, percentage change +7.9%; W=9.0; P<.001; r_b=-0.91, 95% CI -0.96 to -0.78; hypothesis 1), mathematics motivation (+1.1, +20.0%; W=15.5; P=.007; \( r_{bb}=-0.77, 95\%\text{ CI }-0.91\text{ to }-0.43 \); hypothesis 3), and mathematics concentration (+0.8, +12.7%; W=21.5; P=.01; \( r_{bb}=-0.68, 95\%\text{ CI }-0.88\text{ to }-0.27 \); hypothesis 5) scores.

Regarding French classes, even though they were not targeted by the intervention), grade 5 participants demonstrated an increase in motivation (+0.8, +17.7%; W=60.0; P=.05; \( r_{bb}=-0.48, 95\%\text{ CI }-0.76\text{ to }-0.03 \) and self-efficacy (+1.5, +29.4%; W=45.0; P=.02; \( r_{bb}=-0.57, 95\%\text{ CI }-0.81\text{ to }-0.14 \)) scores after the intervention. They also tended to show an increase in academic achievement scores in French after the intervention (+7.3, +11.4%; W=53.0; P=.09; \( r_{bb}=-0.44, 95\%\text{ CI }-0.75\text{ to }-0.03 \). Contrary to hypothesis 1, grade 5 participants showed a decrease in academic achievement scores in mathematics after the intervention (+10.7, +13.6%; W=192.0; P=.006; \( r_{bb}=0.66, 95\%\text{ CI }0.29\text{ to }0.85 \).

Effects on Grade 2 to 4 Participants

The other grades also benefited from the intervention but to a lesser extent. After the intervention, grade 2 participants showed an increase in PL scores (difference +5.2, percentage change +7.1%; W=28.0; P=.03; \( r_{bb}=-0.58, 95\%\text{ CI }-0.84\text{ to }-0.11; \) hypothesis 1) and a marginal increase in academic performance in mathematics (+15.1, +33.7%; W=18.5; P=.06; \( r_{bb}=-0.59, 95\%\text{ CI }-0.86\text{ to }-0.06 \); hypothesis 2). Grade 3 participants showed an increase in mathematics concentration (+0.5, +7.5%; W=50.5; P=.04; \( r_{bb}=-0.51, 95\%\text{ CI }-0.79\text{ to }-0.07 \); hypothesis 5) after the intervention. Regarding French classes, grade 2 participants showed a marginal increase in self-efficacy scores (+1.0, +14.4%; W=24.0; P=.07; \( r_{bb}=-0.54, 95\%\text{ CI }-0.83\text{ to }-0.01 \), grade 3 participants showed a significant increase in academic performance (+16.8, +40.1%; W=68.0; P=.01; \( r_{bb}=-0.58, 95\%\text{ CI }-0.80\text{ to }-0.21 \)) and grade 4 participants showed a marginal increase in concentration (+1.1, +20.7%; W=25.0; P=.09; \( r_{bb}=-0.52, 95\%\text{ CI }-0.82\text{ to }0.01 \)) after the intervention. Grade 4 participants showed a decrease in academic performance in French after the intervention (-15.0, -19.8%; W=103.5; P=.01; \( r_{bb}=0.72, 95\%\text{ CI }0.33\text{ to }0.90 \).
the maximum motivation score before the start of the program. The last teacher rated the students’ motivation at 9.0 before and after the program.

**Discussion**

**Principal Findings**
Since September 2022, primary schools have been required to provide 30 minutes of DPA. In this pilot study, the implementation of an exergaming program as part of the 30 minutes of DPA policy was welcomed by the school’s administration, parents, and teachers, with an increase in perceived motivation for mathematics.

After the program, we observed that children showed increased scores in PL and motivation in mathematics following 11 learning-based exergaming sessions.

**Exergaming Implementation**
Although this PA reform is recent in France, it has already been introduced in other countries several years ago. Indeed, the DPA school policy has been implemented since 2005 in Canada to promote active lifestyles for children in school settings [41], and in the province of Ontario, all elementary school children perform DPA during instructional time [42]. Yet, 10 years later (in 2015), it was revealed that only half of Ontario teachers were meeting this expectation [43], and this number dropped to 23% 5 years later in the report by Martyn et al [44]. The Canadian experience underscores the need to explore effective and sustainable methods for implementing the 30 minutes of DPA in schools. Consequently, this pilot study shows that exergaming can be used as a valuable tool in the deployment of DPA at schools.

**Efficiency and Usefulness of Exergaming**
Our second objective was to find out whether an exergaming intervention could be effective and useful. We hypothesized that implementing a DPA program involving exergaming on a specific academic course could have an impact on different aspects of a child’s experience. First, in line with hypothesis 1, our results showed a significant increase in the PL of the entire cohort, with significant increases of over 7% for both grades 2 and 5. This result is all the more important as it has been highlighted that elevated PL leads to greater PA participation, resulting in positive physiological, social, and psychosocial adaptations, and thus improved physical, mental, and social health [45]. In other words, PL could play a role across the lifespan in promoting positive health. Therefore, exergaming seems to be an effective and useful instrument to promote PL. This observation is in line with the review by Sun [46], which highlighted that active video gaming could contribute to enhancing children’s PL, in particular on the motivational aspect of exergaming, making it possible to provide a variety of opportunities to develop or reinforce basic motor skills among children.

Second, concerning motivational aspects, in line with hypothesis 3, our results showed an overall positive effect on students’ motivation toward the discipline. Indeed, we found an increase in motivation for mathematics (target subject), with a significant increase of almost 10% for the total cohort, while motivation for French (control subject) was not impacted. It seems that allowing students to work on mathematics more entertainingly (ie, by throwing balls onto calculation operations) helps to increase their appeal in this course.

Third, contrary to hypotheses 2, 4, and 5, our results showed no increase in academic performance, motivation, and sense of efficacy in mathematics, but they showed an increase in these variables in French, even though this subject was not directly targeted in the sessions. Although the interpretation is limited without a control group that did not benefit from the intervention, it is conceivable that the participants benefited from additional motivational resources provided by the 30 minutes of DPA toward learning at school, in accordance with the results of Vazou et al [16] regarding motivation and self-efficacy. An argument in support of this explanation may be the marginal increase in general concentration in class for the total cohort, as has been observed in the review by Taras [47], which noted an immediate increase in concentration in students after PA. This overall concentration may have benefited all subjects, especially those frequently considered less difficult than mathematics (ie, French).

Finally, it is important to note that the positive effects of the intervention were found in all school grades, even if a greater benefit was observed in grade 5. As the ability to apply skills or knowledge learned during one activity to another activity is evidence of a transfer process, older children are likely to be more sensitive to it [48]. Indeed, in this study’s intervention, the children were learning with different tasks and objectives (in DPA exergaming and their normal lessons). The transfer of skills from one to the other was therefore not obvious (even if the “mathematics” cue was common to both) and remains a particularly demanding cognitive process for which the children need to be motivated. Decreases in academic performance in French and mathematics (grades 4 and 5) may be explained by constraints in the classrooms, as the teachers were rotated during the semester and the last data collection took place the day before the vacation (the participants were less involved overall in the academic exercises required). However, the marginal increase in mathematics performance in grade 2 and the increase in French performance (overall cohort and grade 3) demonstrate the importance of continuing to test this intervention.

**Practical Implications in the Educational Context**
Learning-based exergames can be powerful allies in the implementation of the DPA policy at schools. For schools and educational teams, the first obstacle could be the associated cost. In France, the Ministry of Education has launched a call for projects entitled, “Pour un socle numérique dans les écoles élémentaires” (“for a digital base in elementary schools”) [49] to equip the schools of tomorrow. In this context, it is necessary to create links between the worlds of research and education. Researchers need to present teachers with the advantages (ie, academic performance, self-efficacy, motivation, PL, PA, and sedentary behavior) and constraints (ie, update, group management, and security) of this type of practice to make the teachers as efficient as possible in different teaching situations. Indeed, as part of the 30 minutes of DPA policy in elementary
schools, one of the major difficulties is sustaining the actions and motivation of teachers, as presented in the Canadian study [43,44]. Once the equipment has been acquired and installed in a fixed position (ceiling-mounted model), one of the solutions for maintaining motivation among teaching teams would be to integrate PA professionals into the internship framework. This option enables teachers to not only benefit from the specific skills of the trainees but also position themselves as observers of the class, to be able to work on specific notions during PE teaching [18].

**Limitations**

In the context of this pilot study, which focused mainly on implementation feasibility and learning, it would have been interesting to consider the children’s physical fitness (ie, muscular strength, agility, and cardiorespiratory fitness) and general state of health. Indeed, a French longitudinal study with a 3-year follow-up of children aged 7.7 years at the start of the study showed that the physical fitness of French youth decreased between childhood and early adolescence [50].

It would also have been interesting to compare the effects of this program with a control group. For example, it would have been worthwhile to compare the scores of the experimental group involving exergaming and targeted school exercises to 2 control groups: the first one with no PA and no school exercises, and the second one with no PA but with school exercises identical to the experimental group (eg, on a tablet computer). To verify the validity of the results, it would also have been necessary to vary the targeted school exercises (eg, mathematics and French; randomizing their inclusion in the intervention to ensure that the most difficult material is not the only one tested). Moreover, this study was carried out in a single school with a single class per level. It would be worthwhile to increase the size of the cohort by increasing the number of classes per level in different schools. Furthermore, our program had a limited duration (3 weeks), and a longer program (at least 1 trimester) with more DPA sessions would undoubtedly have increased the effects we observed and allowed additional benefits to be observed. In future studies, it would be interesting to compare the effects of a short program like ours with those of a longer program, and this could shed light on the duration of effects through time (in particular, following longer interventions).

**Perspectives**

Future studies could explore the possible diffusion effects of enhanced DPA interventions with or without exergaming on various PA indicators (eg, physical fitness, increased mobility by accelerometry, sedentary time and breaks, and increased implication in PE curricula at school). It would also be useful to conduct a longitudinal study to measure the impact of exergaming on not only PA and fitness levels but also the evolution of overweight and obesity in children.

Furthermore, in future studies, it could be relevant to assess intervention effects on students’ academic performance, motivation, and self-efficacy in specific academic courses during interventions, or general attitudes and performances in different courses. Finally, specifying various student profiles concerning these measures (eg, depending on the initial levels of PL and PA, and depending on age or grade) could provide information on the subgroups of children benefiting the most from such exergaming interventions. In addition to student characteristics, it might be useful to consider teacher characteristics (eg, attitudes toward exergaming) to better understand the individual and environmental factors likely to moderate the effects of such interventions.

**Conclusions**

As part of the 30-minute implementation of DPA, the use of learning-based exergaming showed very interesting results in increasing PL as well as student motivation toward mathematics. Furthermore, supporting pedagogical teams with qualified teachers in PA has been proven to be beneficial for both students and staff.

With this encouraging pilot study, it is necessary to continue investigations by increasing the number of students per grade and to carry out research over a longer school period with a control group to confirm these results regarding the use of exergaming on not only PA and fitness levels but also the general state of health. Indeed, a French longitudinal study with a 3-year follow-up of children aged 7.7 years at the start of the study showed that the physical fitness of French youth decreased between childhood and early adolescence [50].

**Acknowledgments**

We would like to thank the school (administration and teaching staff) for allowing us to carry out this pilot study.

**Authors’ Contributions**

AG conceptualized and supervised the project, and contributed to the data collection and processing, project administration, and writing of the original draft. FL, GB, EVC, and EC conceptualized and supervised the project and were involved in reviewing and editing the manuscript. FL contributed to the statistical analysis. AP was involved in daily physical activity implementation supervision and physical literacy during the session. ED and NAB were students in adapted physical activity who carried out the daily physical activity sessions as part of their end-of-year internship.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

PlaySELF adaptation questionnaire.
References


34. Lü Interactive Playground. URL: https://play-lu.com/ [accessed 2023-11-02]

35. PLAYself - Physical Literacy Assessment for Youth. Sport for Life Society. URL: https://playself.ca/playtools/playself/ [accessed 2023-07-18]


37. PLAY Tools. Physical Literacy. URL: https://physicalliteracy.ca/play-tools/ [accessed 2023-11-03]

38. Les outils PLAY. Littératie Physique. URL: https://litteratiephysique.ca/les-outils-play/ [accessed 2023-11-03]


41. PlaySport. Ophea. URL: https://ophea.net/playsport [accessed 2022-04-08]


Abbreviations

DPA: daily physical activity
PA: physical activity
PE: physical education
PL: physical literacy

©Aurélie Goncalves, Florence Lespiau, Gaëtan Briet, Eugénie Vaillant-Coiudard, Angèle Palermo, Elsa Decobert, Nathan Allegret-Bourdon, Elodie Charbonnier. Originally published in JMIR Serious Games (https://games.jmir.org), 23.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
Design of Virtual Reality Exergames for Upper Limb Stroke Rehabilitation Following Iterative Design Methods: Usability Study

Julian Felipe Villada Castillo1, MSc; Maria Fernanda Montoya Vega2, MSc; John Edison Muñoz Cardona3, PhD; David Lopez4, MSc; Leonardo Quiñones4, MSc; Oscar Alberto Henao Gallo5, PhD; Jose Fernando Lopez5, MD

1Faculty of Basic Sciences, Department of Physics, Universidad Tecnologica de Pereira, Pereira, Colombia
2Department of Human-Centered Computing, Exertion Games Lab, Monash University, Melbourne, Australia
3Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada
4Engineering Faculty, Universidad Tecnologica de Pereira, Pereira, Colombia
5Medicine Faculty, Universidad Tecnologica de Pereira, Pereira, Colombia

Abstract

Background: Since the early 2000s, there has been a growing interest in using exercise video games (exergames) and virtual reality (VR)–based interventions as innovative methods to enhance physical rehabilitation for individuals with multiple disabilities. Over the past decade, researchers and exercise professionals have focused on developing specialized immersive exercise video games for various populations, including those who have experienced a stroke, revealing tangible benefits for upper limb rehabilitation. However, it is necessary to develop highly engaging, personalized games that can facilitate the creation of experiences aligned with the preferences, motivations, and challenges communicated by people who have had an episode of stroke.

Objective: This study seeks to explore the customization potential of an exergame for individuals who have undergone a stroke, concurrently evaluating its usability as a technological tool in the realm of physical therapy and rehabilitation.

Methods: We introduce a playtest methodology to enhance the design of a VR exergame developed using a user-centered approach for upper limb rehabilitation in stroke survivors. Over 4 playtesting sessions, stroke survivors interacted with initial game versions using VR headsets, providing essential feedback for refining game content and mechanics. Additionally, a pilot study involving 10 stroke survivors collected data through VR-related questionnaires to assess game design aspects such as mechanics, assistance, experience, motion sickness, and immersion.

Results: The playtest methodology was beneficial for improving the exergame to align with user needs, consistently incorporating their perspectives and achieving noteworthy results. The pilot study revealed that users had a positive response. In the first scenario, a carpenter presents a game based on the flexion-extension movement of the elbow; the second scenario includes a tejo game (a traditional Colombian throwing game) designed around game mechanics related to the flexion-extension movement of the shoulder; and in the third scenario, a farmer challenges the player to perform a movement combining elbow flexion and extension with internal and external rotation of the shoulder. These findings suggest the potential of the studied exergame as a tool for the upper limb rehabilitation of individuals who have experienced a stroke.

Conclusions: The inclusion of exergames in rehabilitation for stroke-induced hemiparesis has significantly benefited the recovery process by focusing on essential shoulder and elbow movements. These interactive games play a crucial role in helping users regain mobility and restore practical use of affected limbs. They also serve as valuable data sources for researchers, improving the system’s responsiveness. This iterative approach enhances game design and markedly boosts user satisfaction, suggesting exergames have promising potential as adjunctive elements in traditional therapeutic approaches.

https://games.jmir.org/2024/1/e48900
Introduction

Background

In recent years, technological advances have influenced motor rehabilitation interventions for survivors of stroke, with the introduction of exergames, known as “serious games for health,” which help motivate individuals in their rehabilitation [1-4]. However, the development of such exergames needs to consider users’ needs and rehabilitation goals [5].

Virtual reality (VR) immersive [6] systems have become increasingly popular in rehabilitation, as they offer immersive and engaging activities, improving motivation and skill acquisition [7]. Nevertheless, systematic reviews have noted that most VR apps primarily focus on balance and gait, with limited attention to upper extremity rehabilitation [8,9].

Efforts have been made to design exergames tailored for survivors of stroke, but challenges remain, including limited user involvement and lack of immersive VR integration [10,11].

This study aims to address these challenges by designing a VR-based upper limb rehabilitation exergame using a user-centered approach, involving survivors of stroke in the design process and conducting playtests with an immersive VR setup [12]. The methodology aims to improve interdisciplinary collaboration and facilitate the involvement of clinicians in the design process [13,14]. The primary objectives are to provide personalized upper arm physiotherapy for survivors of stroke through an improved VR exergame and to assess its usability through user feedback [15]. This work encourages collaboration among clinicians, researchers, and designers to create an engaging rehabilitation exercise that complements the recovery process for survivors of stroke, ultimately enhancing their quality of life.

Related Work

VR-Based Physical Rehabilitation for Stroke

Experts in rehabilitation, kinesiology, and neuroscience are integrating VR systems with exergames to enhance the appeal and effectiveness of rehabilitation processes [16]. Early studies, such as those by Henrique et al [17] and Burke et al [18], demonstrated the positive impact of exergames on balance, gait, and upper limb motor function in patients with stroke, highlighting improved therapy adherence [14,16-20]. However, systematic reviews have indicated that most VR apps for after-stroke therapy primarily focus on balance and gait, with limited attention to upper extremity rehabilitation [3,21,22]. To address this gap, we aim to evaluate the potential of an exergame for upper limb rehabilitation using immersive VR systems [22].

In addition, prior research has shown that complementing or replacing standard rehabilitation with VR-based rehabilitation can result in significant improvements in gait speed, balance, and mobility in patients with stroke [3,17,21-23]. Our work aims to contribute to the development of guidelines for using VR-based rehabilitation in conjunction with conventional therapy, with a focus on upper limb rehabilitation.

Although some researchers, such as Reis et al [10], Leung et al [11], and Horsham et al [24], have proposed methodologies for developing specific exergames for stroke rehabilitation, there is still limited knowledge regarding immersive VR-based designs targeting upper limb rehabilitation [10,11,24]. Therefore, we intend to involve survivors of stroke in an iterative playtesting process to develop an upper limb VR-based rehabilitation system and bridge this gap.

Playtesting as an Iterative Design for Stroke

This section covers research related to the use of playtesting as an iterative user-centered design (UCD) methodology. UCD has played a significant role in the development of games for rehabilitation and overall health [25,26], as it is a methodology that allows active participation of the target population in the system’s prototyping process. UCD, applied in game design, often advocates for an interactive and participative methodology that includes multiple playtests with end players. Playtesting is an activity carried out with potential users or players who interact with game prototypes developed in the early stages, making it easier to gather individual opinions and ideas that contribute to improving the gameplay aspects of exergames during their development [27,28]. Playtesting is a key and standardized methodology used in game studios to iterate and systematically improve games before they are released to the public [29].

A relevant example is the work of Duval et al [30], who conducted a collaborative study with 14 clinicians, focusing on therapeutically validating the game based on their opinions rather than those of users. Duval et al [30] obtained significant findings by addressing the adoption of therapy and personalizing it according to the characteristics valued by medical professionals. In contrast, other UCD works, such as the study by Aguilar et al [31], have not used playtesting but have used usability tests involving scales and flow state questionnaires. Findings from 3 years of experience with exergames developed for older adults using UCD methods concluded that devoting the key to engaging with end users and considering feedback and opinions can be considered the best practice guide for the development of therapeutic games [32]. We believe that playtesting can be beneficial for the design process of games for health, as it is strongly recommended to involve the target audience during the game design and development processes. By doing so, developers increase the likelihood of creating games that consider the specific preferences, motives, and characteristics of survivors of stroke in need of physical therapy [21]. By including survivors of stroke in interactive playtesting and, consequently, in enhancing a VR exergame, we begin to understand how this design methodology affects the subsequent use of the exergame as a therapeutic tool.
Methods

In this section, we introduce the interdisciplinary team that worked on the improvement of the VR exergame we used in the playtesting session and the pilot study, as well as the description of the VR exergame. Furthermore, we present the playtesting methodology and the pilot study methodology.

Interdisciplinary Team

The structured design team was composed of an expert clinical physiatrist who advised the movements that users with stroke are likely to perform from a clinical viewpoint; a physiotherapist who provided permanent follow-up in all sessions with the users; a designer of exergames who helped implement the UCD methodology to have clear game mechanics; an expert in biomechanics who analyzed ranges of movement, postures, and gestures; a user experience researcher who organized all sessions with the users; 2 professional game programmers who created the game prototypes; and 2 users who experienced stroke episodes and interacted with the system and based on the answers they gave us an improved exergame. For 2 months, this group convened weekly to discuss the exergames’ requirements, technologies, and overall scope of the project. The discussions were centered on defining the activities in the internet-based environment and strategies for the recruitment of potential users. At the end of the design process, the group of game developers with programming experience used the Unity game engine (Unity Technologies) to materialize the ideas.

The main topics addressed by the interdisciplinary team were (1) the definition of the main objectives and roles of the project; for example, project management was assumed by the exergames designer, and the user experience researcher assumed the role of project manager and conducted most of the fieldwork; (2) socialization of playtesting activities, including user recruitment and experimental protocol; and (3) reconsideration and further adjustment of game design elements, such as game mechanics and their mapping with therapeutic objectives.

Design of the VR Exergame

Prior Design of the VR Exergame

We performed a rapid contextual design based on a previous study, where user profiles were defined using user personas [33]. In this study, we characterized 4 persona roles that distinguish them as gamers: apathetic, empathetic, beginner, and experienced. Specifically, we used the results of the user modeling process to define certain game elements. For example, we found that users showed interest in sports games. Hence, body interaction familiar with certain sports games was an important requirement to be integrated into the exergames. In addition, users were comfortable with game content related to their daily lives. Therefore, incorporating cultural activities into the exergames could be a promising approach. In addition, in this prior study, we considered the following clinical requirements when developing the VR exergame [34,35].

Population specificity: according to previous studies [36], most users who have experienced a stroke are older than 40 years, and few are younger than 30. These studies also showed that the older population has little experience with VR. In contrast, therapist experience suggested that ranges of motion vary across users who have experienced a stroke. Therefore, designers should be careful when adapting internet-based therapy to a wide range of capabilities [37].

Motor learning: different principles of motor recovery should be considered in the creation of the activities to be performed within the exergames, such as a meaningful task, intensive and repetitive practice, movements close to the normal range, muscle activation that drives the practice of movement, and variability and progression of training [10,38,39].

Rehabilitation movements: the movements suggested by clinicians to perform upper limb physical therapy in paretic users are listed in Table 1. These are shoulder flexion and extension, elbow flexion and extension, and Kabat diagonals. Kabat diagonals are internal and external rotation movements of the shoulder. According to Della Tommasina et al [40], a repetitive process of these movements is necessary to perform physiotherapy, from which a more effective range of motion recovery will be obtained [41]. The elbow and shoulder are the upper limbs’ main joints, articulating the arm’s largest segments. These joints require a greater range of motion in flexion and extension and often affect and limit arm motion when a stroke episode occurs. Hence, we decided that users in a seated position should perform different arm movements while playing the VR-based rehabilitation exergame, targeting multiple possible physical rehabilitation needs in the upper limbs of people with stroke.

Table 1. Rehabilitation movements proposed for the exergame.

<table>
<thead>
<tr>
<th>Movement</th>
<th>Application</th>
<th>Action in the exergame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elbow</td>
<td>Flexion and extension</td>
<td>Improvement in the width of movement in daily life activities</td>
</tr>
<tr>
<td>Shoulder</td>
<td>Flexion and extension</td>
<td>Improvement in the width of movement in daily life activities</td>
</tr>
<tr>
<td>Elbow</td>
<td>Interior and exterior extension</td>
<td>Improved range of motion in the daily life activities</td>
</tr>
</tbody>
</table>

Complementary therapy: we propose a therapy that uses the VR-based rehabilitation exergame to complement the rehabilitation process instead of replacing the traditional one, such as the one proposed by Goncalves et al [42]. We are confident that users who have experienced a stroke will play an exergame with engaging activities because they are developed based on their needs and motivations. Therefore, the exergames will allow a disruptive experience different from conventional therapies, generating interest, excitement, and willingness to carry out their rehabilitation process without neglecting the conventional therapy recommended by clinical specialists [38]. Considering the rapid contextual design and the clinical
requirements named in this study, in a prior study, we developed a VR exergame following the well-known game design methodologies by Schell [29], as we contextualize in the following subsection.

**Motion Health VR: VR Exergame for Stroke Rehabilitation**

In a prior study, we established a game design concept and game mechanics using the methodology proposed by Schell [29] and a complete contextual design conducted with users with stroke. This study covers a more systematic and complete description of the playtesting sessions conducted with players with stroke to iterate and improve the game based on the initial concept. Knowing the preferences, ages, and profiles of potential users, we decided to explore a design concept for the cultural regions of Colombia (the Caribbean, Pacific, Andean, Orinoco, and Amazon). We discovered that older adults are inclined to engage in activities in the countryside and typical and authentic Colombian games. We discussed this exergame concept with the clinicians of our design team, gathering feedback regarding the potential movements to be performed, particularly in the context of mapping them for stroke rehabilitation therapies. Considering the scope of the project and previous research conducted with local users, development capabilities, and timelines, we decided to start by developing a game design concept related to the activities of the Colombian Andean region. The Andean region is the central region of Colombia and has crosscutting activities that are representative of the entire country and run throughout the central Andes. The population of this region practices sports, such as *sapo* and *tejo* (throwing games), rowing, and other more well-known sports, such as basketball and boxing. They also engage in other daily activities, such as fruit picking, horseback riding, and bush cutting [43].

Aligned with these cultural activities, we designed game scenarios that focused on a local setting using Colombian games. The design team analyzed existing VR games to establish game mechanics that could involve desired rehabilitation movements, always considering the player’s motivators and needs. This analysis facilitated communication between the team of clinicians and specialists and the design and development team while also helping to specify the activities that would be familiar and engaging to the users. Therefore, we called the VR exergame “Motion Health VR.” The exergame comprises 3 main scenes in which players must develop 3 different activities: hammering, throwing a metal disk (a traditional Colombian game called *tejo*), and cutting bushes while riding a horse.

The exergame presents 3 meticulously crafted scenarios, each aligned with its unique reference to Figure 1. In the “carpenter” scenario (Figure 1A), players engage in a dynamic elbow flexion and extension challenge, wielding a hammer to systematically crush boxes that vary in color and size, demanding specific ranges of motion. Players must skillfully adjust their proximity to the boxes to adapt to this diverse challenge, seamlessly weaving in back-and-forth and crossbody-reaching movements. Between the box-smashing activity, a captivating puzzle gradually unveils itself, featuring distinct Andean wildlife. Upon completing the activity, players earn the gratifying experience of visualizing the completed animal puzzle. As players enhance their hammering skills, the game dynamically escalates in difficulty either by increasing hammering frequency or by reducing box sizes. In the second scenario, inspired by Colombia’s traditional *tejo* game (Figure 1B), players embark on a shoulder-focused flexion and extension adventure, mirroring the popular sport played nationwide. Throwing a metal disk toward an explosive target known as a *mecha* on a clay court, players must adjust their shoulder movements according to the target’s distance, finetuning their range of motion for precise throws and aiming to maximize target hits with minimal repetitions. In the third scenario, the “farmer” (Figure 1C), players are challenged with a multifaceted movement that combines elbow flexion, extension, and internal and external rotation of the shoulder, akin to Kabat diagonals. In this rural setting, players ride an internet-based horse while wielding a machete, a staple tool in the Colombian countryside, tasked with clearing the obstructive bushes that appear on both sides of the road. With one arm gripping the machete and the other resting on the horse’s rein, players face escalating challenges as the game progresses.

**Figure 1.** The presented scenarios of the Motion Health VR exergame with (A) a carpenter, (B) a throwing activity, and (C) a farmer, based on the movement of the Kabat diagonals.

**Playtesting Sessions**

Iterative game design involves playtesting sessions with end users, who will shape the game features before its final deployment. The main objective of playtesting sessions is to iterate different playable prototypes to improve the overall playability of the game, thus increasing the likelihood of adoption. In addition, playtesting allows the researcher to assess the ability of potential players to perform the proposed activities and understand game feedback. We developed playtesting sessions using the VR exergame designed in a prior work.
The playtest sessions were guided by a researcher accompanied by a clinical specialist who helped contact different users who had experienced a stroke. Owing to the COVID-19 pandemic, visits to each user were scheduled in such a way that biosafety protocols were maintained (eg, distancing, constant use of masks, and hand and footwear disinfection). Upon arrival at the agreed location, stakeholders performed the recommended distancing protocols. Then, the researcher prepared the experimental protocol, which consisted of setting up a table, a chair without a hand rest, and verifying the internet connection. Prototypes of the exergames were developed before the playtesting sessions and ported to the VR headsets (Oculus Rift in the first 4 iterations and Oculus Quest in the final version). Disposable headset protectors and cleaners were used to maintain biosafety measures. Table 2 presents the structure of the playtest sessions.

<table>
<thead>
<tr>
<th>Actions</th>
<th>Considerations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brief introduction of the dynamics of the session and interaction between the researcher and the user with stroke to obtain informed consent</td>
<td>a</td>
<td>10</td>
</tr>
<tr>
<td>A quick explanation of how the VR system works and what to expect from the activity</td>
<td>Preparation and arrangement</td>
<td>5</td>
</tr>
<tr>
<td>System implementation (HMD, headphones, and controls)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Free play or natural interaction with the system</td>
<td>Manifestation of difficulties, in real time if necessary</td>
<td>5</td>
</tr>
<tr>
<td>Receive feedback or explore ideas while users play</td>
<td>Formulate the questions established for the session</td>
<td>10</td>
</tr>
<tr>
<td>Conclude the session with questions about the experience</td>
<td>Questions</td>
<td>10</td>
</tr>
</tbody>
</table>

aNot available.
bVR: virtual reality.
cHMD: head-mounted display.

A total of 4 playtesting sessions involving 9 end users who had experienced a stroke were conducted. Each session was performed with a minimum of 2 users chosen considering their availability (Multimedia Appendix 1 lists the users who participated in each session and their demographic information). After playtesting, we recorded a video summarizing the session and documented a brief analysis that was subsequently discussed with the research and design teams. After each playtest, the team held a general meeting where all the discussions were presented. The subsequent playtest was scheduled after the implementation of the suggested game changes.

Pilot Study: Evaluation of the Game Experience and Usability of the Exergames

After conducting game playtests and completing a playable prototype of the VR exergame (4 iterations), we decided to carry out a pilot study to evaluate the usability of the game with a group of users who had experienced a stroke, in which 2 users who participated in one of the 4 iterations were part of the pilot study group.

We conducted a 20-minute session that was part of the rehabilitation therapy in which users who had experienced a stroke played the iterated version of the Motion Health VR exergame. We ported the final version of the game to the standalone VR headset, Oculus Quest 2, as it has several advantages, such as being wireless, comfortable, and having a high image resolution.

Users

This usability study was developed with 10 users who had experienced a stroke contacted through the clinician and therapist of the design team. We chose this sample size conveniently, considering the availability of users, which was very limited. In contrast, the small sample size allowed us to follow the biosafety protocols required for the COVID-19 pandemic, which was still ongoing in Colombia at the time of the study. The inclusion criteria for the study were being aged >50 years, having experienced a stroke and having hemiparesis or monoparesis, being able to read and write, not having serious vision problems (eg, strabismus), and not having diagnosed cognitive disabilities (eg, dementia).

Ethical Considerations

The bioethics committee of the local university approved this study, which was also approved by the bioethics committee of a local rehabilitation center (52–050623). Users volunteered for this study and agreed to participate by signing an informed consent form.

Usability Study

Two questionnaires (instruments) were used to assess the game user experience immediately after interacting with the immersive game.

Virtual Reality Neuroscience Questionnaire

The Virtual Reality Neuroscience Questionnaire (VRNQ) measures the quality of user experience, game mechanics, and in-game assistance. It comprises 20 questions, each scored on a Likert-type scale ranging from 0 to 5 [44]. The advantage of using this questionnaire is that it provides the limits to assess the suitability of the software in VR [44]. VRNQ produces a total score that reflects the overall quality of the VR software and 4 categories as follows: (1) game experience, where the level of immersion and pleasure of the experience are evaluated; (2) game mechanics, where user interaction in the internet-based
The Immersive Tendencies Questionnaire (ITQ) determines the differences in an individual’s tendencies to experience immersion and presence after interacting with a VR scenario. ITQ comprises 18 questions rated on a Likert scale from 1 to 7, resulting in a possible score ranging from 18 to 126. In the original study, the mean score of the samples was 76.66 [45]. This questionnaire is considered a standard in VR research and has been widely used in different applications [22]. A user with a positive immersive tendency based on the ITQ score is likely to experience higher levels of VR presence, which has been associated with better task performance [22,41,46]. This questionnaire is useful because it allows the evaluation of immersion in a way that does not depend on the specific internet-based environment, making it possible to determine independently if an internet-based environment performs poorly or if the statistical sample has low immersion trends. Some ITQ questions are as follows:

- Does it often happen that while daydreaming, you forget what is happening around you?
- Does it happen to you that you are so engrossed in a movie that you forget what is happening around you?
- Do you identify with television characters?
- When you use an exergame, does it occur to you that you feel like you are inside the game instead of sitting down using the controller?
- Do you stay scared for a while after watching a scary movie?

Experimental Setup

The researcher and the physical therapist held the interaction session at each participant’s home, where they chose a comfortable space to set up the VR system. The setup consisted of a chair in which the user with stroke was seated with the Oculus Quest 2 wearable headset and its respective wireless controllers. The researchers were able to see what the players were doing via the official Meta Quest app using an electronic tablet in real time.

Protocol

The users began the pilot study session seated, using the headset and holding the controllers. We conducted the session in the following order. The user who had experienced a stroke performed an upper body warm up for 5 minutes, guided by the physical therapist. The researcher and physical therapist prepared the user for the game by helping them put on the VR system. The researcher started the exergame, which was presented throughout the session, accompanied by a physiotherapist. After the interaction, the user who had experienced a stroke completed the 2 proposed questionnaires. Given the biosafety regulations established by the Colombian government because of the COVID-19 pandemic, all those involved in the sessions always wore masks. In addition, the researcher cleaned all VR and mounting elements each time they were used.

Data Analysis

The questionnaires were scored following the instructions of previous works. Descriptive statistics, such as mean and SD, were calculated and reported [47,48].

Results

Overview

The results are presented in 2 subsections. The first subsection details the transformation process of the game after conducting 3 playtesting sessions and iterations involving the end users and the interdisciplinary design team. The second subsection presents the preliminary results of a pilot study evaluating the user experience of the final game with a group of 10 players with stroke.

Playtesting

The following subsections detail the results of the iterative design process of the Motion Health VR exergame, reporting the details of the playtesting sessions and the modifications made to each iteration. The overall objectives of the playtest were to validate the acceptance and playability of users with stroke and to explore whether they were able to perform the activities proposed in the VR scenarios and game mechanics. In addition, analysis of errors and optimization of game mechanics were crucial to improve playability. A total of 9 users with stroke were involved throughout the 4 playtesting sessions conducted (Multimedia Appendix 2), focusing on certain game elements (eg, mechanics and esthetics) via playable prototypes and reporting back to the design team.

Analysis of Playtest 1

This version of the exergame was created to test the first 2 scenarios. Players 1 and 2 (U1 and U2) participated in this session following the protocol in Table 2. We used a VR-ready laptop and an Oculus Rift headset with controllers. The playtesting goals were to evaluate the appropriateness of the proposed range of movement for hammering and throwing the disk and to explore button combinations for performing the activity using the controllers. We found that (1) the game should consider different scales of spasticity to provide a more adaptive experience [49]; (2) the buttons should be suspended from their functions to avoid triggering involuntary functions; and (3) a rest period should be granted to the user because, as recommended by the physiotherapist, long periods of exercise generate symptoms of fatigue.

Analysis of Playtest 2

The objective of these playtests was to test the modifications introduced in the first 2 scenarios based on the considerations in the Analysis of Playtest 1 subsection. Players U3, U4, and U5 participated in this playtest following the protocol in Table 2. A VR-ready laptop and an Oculus Rift headset with controllers were used. Figure 2 shows the evolution of the 2 scenarios after the first playtest, showing the improvement in
content according to the real scenarios where these activities were performed.

**Figure 2.** The final version of each scenario of the Motion Health VR exergame: (A), the carpenter scenario, (B) the tejo scenario, and (C) the farmer scenario.

Users with stroke reported an improvement in the simplicity of the interaction, as they found it much easier to perform the movement owing to an initial calibration of the position added to the 2 scenarios, which adjusts the player’s position concerning the internet-based surroundings, ensuring objects are at a reachable distance. In addition, we found that (1) the objects in each scenario should be in a static position; thus, people can avoid unnecessary displacements within the internet-based space that can generate dizziness; (2) we need to improve the auditory feedback of scenarios to create an immersive experience; and (3) we need to improve the calibration scene to allow players with low mobility to perform the tasks.

**Analysis of Playtest 3**

The third scenario was prototyped, and the game mechanics were ready for playtesting. The objective of this playtest was to test whether players could easily understand and interact in the farmer scenario by performing the proposed movements, that is, riding the horse while holding the rein with the unaffected arm and cutting bunches with the machete using the affected arm. Players U6 and U7 (Multimedia Appendix 3) participated in this test following the protocol in Table 2. We used a VR-ready laptop and an Oculus Rift headset with controllers. We found that (1) it would be useful to place the avatar on the horse from the beginning and (2) the game should allow cutting bushes to be performed using both arms and provide adequate time to switch the game controller between hands because the players become tired after certain repetitions.

**Analysis of Playtest 4**

After iterating each scenario and exploring potential pitfalls and interaction errors, a final playtesting session was scheduled to test the overall functioning of the integrated system. This prototype of the exergame presented an embellishment of the contents (Figure 2), which was an improvement in the overall aesthetics of the game. In addition, a structured exercise session was recommended by the clinical rehabilitation experts, following a 15-minute session (similar to other studies of the same nature [50]). Therefore, the 3 scenarios were presented in sequence and switched after 5 minutes (approximately). Players U3, U4, and U5 participated in this test following the protocol in Table 2. We used a VR-ready laptop and an Oculus Rift headset with controllers.

We found that (1) scene transitions should be smoothed and (2) the game should implement rest periods between each scenario, as users still manifested mild fatigue from performing so many repetitions while preparing for the next mechanic.

Finally, we integrated the above recommendations into the scenarios and developed the final prototype of the Motion Health VR exergame.

**Evaluation of Game User Experience**

This section presents the results of evaluating the game user experience of the co-designed Motion Health VR exergame involving 10 users with stroke (Multimedia Appendix 4). For this part, the game was modified to a more portable, standalone, and easy-to-use headset, the Oculus Quest 2. Only the final deployment platform was changed (from wired to wireless VR), and no other changes were made. The questionnaires were administered at the end of the session, asking users to rate their experience in a wide range of aspects following the ITQ and VRNQ.

**Virtual Reality Neuroscientific Questionnaire**

The results of administering VRNQ are reported as average values with SD (Table 3). Each category had a maximum of 35 points. Gaming experience was rated at a mean of 24.8 (SD 4.5), game mechanics mean 23.8 (SD 5.5), game assistance mean 23.9 (SD 5.5), and motion sickness (inversely proportional), mean 31.1 (SD 5.6). Consequently, the maximum possible general score for this test was 140, in which the exergame Motion Health VR obtained a mean of 103.6 (SD 19.4). This level of quality is considered more than adequate as it exceeded 100 points. From this, it can be concluded that the users experienced a high level of immersion during their video game experience, and the quality of the Motion Health VR exergame obtained a general average of mean 103.6 (SD 19.4), which is considered an adequate quality because it exceeded 100 points [47]. On the basis of this result, we observed that users had a high immersion index; the experience with the exergames was very pleasant; and the quality of the graphics, sound, and technology, in general, was perceived as very positive. Finally, the system showed the best results in the motion sickness index, which shows that exergames did not cause major side effects associated with cybersickness or nausea [47,51].

https://games.jmir.org/2024/1/e48900

JMIR Serious Games 2024 | vol. 12 | e48900 | p.113
Table 3. Virtual Reality Neuroscientific Questionnaire (VRNQ) categories.

<table>
<thead>
<tr>
<th>VRNQ categories</th>
<th>Score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game experience</td>
<td>24.8 (4.5)</td>
</tr>
<tr>
<td>Game mechanics</td>
<td>23.8 (5.5)</td>
</tr>
<tr>
<td>Game attendance</td>
<td>23.9 (5.5)</td>
</tr>
<tr>
<td>Motion sickness</td>
<td>31.1 (5.6)</td>
</tr>
</tbody>
</table>

**Immersive Tendencies Questionnaire**

ITQ was used to evaluate users’ immersion experience and presence following their engagement with the exercise. The overall ITQ score averaged 60.8 (SD 11.6), signifying that participants with a history of stroke perceived relatively low to moderate levels of immersion and enjoyment (Table 4). In terms of concentration (mean 22.3, SD 2.9, with a maximum score of 35), users consistently achieved high scores, indicating that the game effectively captured their attention.

Regarding immersion (mean 17.9, SD 4.7, with a maximum score of 35), users reported a sense of engagement with the game. In terms of emotions (mean 14.6, SD 5.1, with a maximum score of 28), the findings suggest that users developed a strong emotional connection with the game [33,51,52].

Table 4. Immersive Tendencies Questionnaire (ITQ) categories.

<table>
<thead>
<tr>
<th>ITQ categories</th>
<th>Score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentration</td>
<td>22.3 (2.9)</td>
</tr>
<tr>
<td>Immersion</td>
<td>17.9 (4.7)</td>
</tr>
<tr>
<td>Emotion</td>
<td>14.6 (5.1)</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>6.1 (2.2)</td>
</tr>
</tbody>
</table>

**Discussion**

**Principal Findings**

This study summarizes our efforts to report methodological approaches extensively used in game design that have a significant value when used to design VR exergames for stroke. We also showed the results of a preliminary usability test conducted involving 10 users with stroke who played the Motion Health VR exergame after completing 4 iterations using playtesting sessions. Overall, the VR exergame exhibited medium-to-high levels of game user experience and low levels of perceived symptoms associated with VR, such as nausea and dizziness. Moreover, regarding enjoyment, users expressed a high willingness to participate in therapies and continue with the sessions. The user experience questionnaires showed that users experienced increased immersion, emotional connection, and enjoyment with the VR exergame, although the concentration remained consistent. These results are consistent with previous research using immersive VR in older adults [53] and people with stroke [54,55]. Furthermore, we have carefully reported the methodological aspects related to playtesting sessions with users with stroke and specific procedures to conduct such sessions. From the playtesting sessions, we can extract the value of evaluating prototypes in the early stages of the game design process because this prevents researchers from struggling with complex interactivity and usability issues later during the trials. As reported in the study by Toro et al [56], early involvement of end users in VR systems for exercise promotion is a desired practice, and it is not commonly used among those creating custom-made exergames for older adults [57-59].

**Playtesting as a Tool for Iterative Design**

Playtesting is a part of the iterative design methodology used in different UCD approaches [51,60]. In our case, we performed playtest with several users who had experienced a stroke, which allowed us to improve the content, playability index, and game mechanics from an ergonomic approach to provide greater user comfort. Although prior designs of physical rehabilitation games for stroke have involved UCD [61], involving users with stroke in playtesting and follow-up sessions is not common [50]. We suggest that designers consider including playtesting with users with stroke because, in terms of rehabilitation therapy performance, playtesting revealed important details, such as the importance of performing a calibration stage or removing the buttons and other interactions with the VR equipment. We found that this stage provides exergames with the characteristics to adapt to the physical needs of each user, such as the range of motion and spasticity scale of each user [62]. In the context of serious games, the importance of adaptive games has increased, as every user has different requirements [62]. Furthermore, the playtesting methodology allowed us to strengthen relationships with all stakeholders, from developers to clinicians. This aligns with the findings of previous studies that have emphasized how involving multiple stakeholders in the design process leads to a more suitable and user-centered prototype [63]. The importance of maintaining close relationships with stakeholders has also been underscored, and this study reaffirms the relevance of this challenge in the successful implementation of playtesting. Effective collaboration between developers and clinicians, driven by the willingness and availability of both parties, has been a recurring theme in the literature [62]. Previous studies have pointed out the pivotal
role of this relationship in the success of VR rehabilitation programs following stroke.

In particular, we noticed that playtesting allowed a stronger relationship with patients with stroke, increasing their willingness to participate in future studies. Although other researchers have reported difficulty finding specific populations to be involved in studies [64,65], considering our final results, we highly recommend that research teams plan to conduct multiple playtests before conducting studies.

Finally, we consider that after performing 4 playtests, based on the observations of the users who had experienced a stroke and the recommendations of the clinicians, an optimal version of the exergame was obtained, which could also be used by designers to facilitate piloting and prevent errors during data collection. As mentioned in this study, the inclusion of playtests and the collaboration of specialized professionals in programming and design align with the best practices recommended by previous research [66].

Usability of UCD VR Games in Patients With Stroke

We observed that ITQ and VRNQ scores were below the expected mean, as reported in the Results section. Similar results have been reported previously because there were some concerns about the usability of VR in older adults, including those who have had strokes. A systematic review of clinical and research applications of VR in older people identified usability issues, such as discomfort, cybersickness, and difficulty with the equipment [67]. Therefore, based on the current results, the design team must improve the gameplay mechanics and usability of the VR exergame to achieve better results in gameplay experience metrics before the subsequent trial. Nevertheless, the system scored high in usability, as its overall score was higher than expected (100 points). Furthermore, as our results showed that the VRNQ category with the highest score was motion sickness, meaning that users who experienced a stroke felt little nausea, our work aligns with studies using similar VR apps [8,20,47]. These usability results can guide other exergame designers to adjust their apps to suit older adults.

Use of Interactive Technology for Telehealth Care

The use of portable and autonomous technologies, such as the Oculus Quest 2 headset, during the pandemic has been an innovative response supported by this research [68]. This reinforces the notion that virtualization of health care, driven by technology, is becoming increasingly important. A recent review describing the promising landscape of telerehabilitation tools aided by serious games for upper limb stroke rehabilitation highlights the evidence of efficacy, the need for further research in this area, and the promise of digitally connected games to complement conventional rehabilitation [69]. Nevertheless, although VR has never been more accessible before, the reality is that the cost-effectiveness of its use in telehealth programs in both rural and urban areas in Latin America is still very limited [70]. In summary, our study contributes to emerging efforts in which interdisciplinary collaboration and the use of innovative technology during times of crisis, such as the pandemic, continue to draw a research pathway in this continually evolving field [63,71,72].

Limitations

We developed this study between August 2020 and June 2021, when most rehabilitation centers were closed owing to COVID-19 pandemic restrictions. Therefore, accessing users with stroke was a difficult task that we overcame with the help of the therapists. They provided us with the contact list of their former users, and we contacted them personally. Notably, although we took all safety precautions, people with stroke feared contagion and only a couple of them participated in the playtesting sessions. We acknowledge the lack of homogeneity in the sample of users who participated in the playtesting because stroke is a condition caused by several factors and affects both sexes, and users who have experienced a stroke tend to have a wide range of ages and ethnicities. The small sample size may limit the generalizability of our findings. Moreover, this limitation was difficult to address because of the COVID-19 pandemic restrictions that were under regulation when we developed this study. That is why, for the pilot study, we limited the users’ age to >50 years. VR is a technology that is constantly changing and improving, in the sense that we started playtesting with the Oculus Rift headset and then moved on to using the Oculus Quest 2 for the pilot study because of its portability advantages. We overcame these technological changes because of the cross-platform features offered by the game engine used (Unity).

However, VR content development is a challenge when designing deployable solutions. Nonetheless, despite the pandemic situations in which this study was developed, the iterative design and preliminary study were carried out owing to the implementation of portable and autonomous tools, such as the Oculus Quest 2 VR headset, which are becoming increasingly important in the virtualization of health care delivery.

The duration of the usability study was very short, and users only interacted with the final game in a single session lasting approximately 15 minutes. A short-term study may not adequately capture the long-term benefits or challenges of using VR exergames for stroke rehabilitation. We plan to extend this initial pilot study and conduct a single-arm longitudinal study involving a similar group of users for 12 sessions for 3 months. Furthermore, this study did not include a control group for comparison. Without a control group undergoing traditional rehabilitation methods, it is challenging to conclusively attribute improvements to the VR exergame alone. Finally, although the study emphasizes the iterative design process and user feedback, it is difficult to know how these design changes directly affect the rehabilitation outcomes. Future studies with this game should include rehabilitation outcomes such as upper limb range of motion and spasticity levels.

Conclusions

Our research has conclusively demonstrated that creating VR exergames for stroke rehabilitation by involving end users early in the design stages brings advantages such as reducing interaction errors and unnecessary game design elements that do not contribute to the therapy. UCD is highly recommended as a design methodology for creating games specifically tailored to the rehabilitation of users who have experienced strokes. The
results of this study support the effectiveness of multiple playtesting in producing therapeutic games that align with the needs and abilities of users with stroke, such as creating familiar internet-based environments and activities and removing unnecessary motion that could lead to motion sickness. This conclusion underscores the importance of adopting a patient-centered approach in the development of medical apps and technologies for rehabilitation. Our findings have yielded promising results regarding the use of immersive VR in the context of upper limb stroke rehabilitation. Users who immersed themselves in internet-based environments using custom-built exergames showed good levels of immersion and enjoyment and reduced levels of perceived nausea or dizziness. These results suggest that VR technology holds potential as a therapeutic tool for the treatment of users with stroke-related impairments, especially for at-home therapies. However, further research and long-term follow-up are required to fully understand the scope and limitations of this technology in the rehabilitation of this user group. The use of playtesting as an iterative tool for enhancing video game design enables comprehensive interaction with the user. This interaction allows for genuine customization of the therapy, leading to the development of a video game tailored specifically for users who have experienced stroke.

Acknowledgments
The authors would like to thank Felipe Gomez for his contribution as a physical therapist to the design team. The authors also thank game programmer Ricardo Arango, who contributed to the construction of the exergames. Finally, the authors would like to thank the Comfamiliar Hospital and all the users who helped with testing the exergames.

Data Availability
The data sets generated during or analyzed during this study are available from the corresponding author upon request.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Users in the playtests.
[DOCX File, 15 KB - games_v12i1e48900_app1.docx]

Multimedia Appendix 2
Demographic data of the user's study pilot.
[DOCX File, 15 KB - games_v12i1e48900_app2.docx]

Multimedia Appendix 3
Changes made throughout the iteration process of the playtests.
[DOCX File, 16 KB - games_v12i1e48900_app3.docx]

Multimedia Appendix 4
Scenario sketches.
[DOCX File, 60 KB - games_v12i1e48900_app4.docx]

References


32. Chu CH, Biss RK, Cooper L, Quan AM, Matulis H. Exergaming platform for older adults residing in long-term care homes: user-centered design, development, and usability study. JMRI Serious Games 2021 Mar 09;9(1):e22370 [FREE Full text] [doi: 10.2196/22370] [Medline: 33673327]


64. Saltiveri TG. MPIu+a. Una metodología que integra la ingeniería del software, la interacción persona-ordenador y la accesibilidad en el contexto de equipos de desarrollo multidisciplinarios. Universitat de Lleida. 2007. URL: https://www.tdx.cat/handle/10803/8120#page=1 [accessed 2023-12-11]
69. Villada Castillo et alJMIR SERIOUS GAMES

https://games.jmir.org/2024/1/e48900
JMIR Serious Games 2024 | vol. 12 | e48900 | p.119

(page number not for citation purposes)


Abbreviations

ITQ: Immersive Tendencies Questionnaire
UCD: user-centered design
VR: virtual reality
VRNQ: Virtual Reality Neuroscientific Questionnaire
A Serious Game (“Fight With Virus”) for Preventing COVID-19 Health Rumors: Development and Experimental Study

Shuo Xiong1, PhD; Long Zuo2, PhD; Qiwei Chen3, MA; Zhang Zeliang4, PhD; Mohd Nor Akmal Khalid4,5, PhD

1Philosophy and Social Sciences Laboratory of Big Data and National Communication Strategy, Huazhong University of Science and Technology, Wuhan, China
2School of Information Engineering, Chang’an University, Xi’an, China
3School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan, China
4School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
5School of Computer Science, Universiti Sains Malaysia, Georgetown, Malaysia

Corresponding Author:
Shuo Xiong, PhD
Philosophy and Social Sciences Laboratory of Big Data and National Communication Strategy
Huazhong University of Science and Technology
Building 6, No 1037, Luoyu Road
Wuhan, 430074
China
Phone: 86 15927188806
Email: xiongshuo@hust.edu.cn

Abstract

Background: Health rumors arbitrarily spread in mainstream social media on the internet. Health rumors emerged in China during the outbreak of COVID-19 in early 2020. Many midelders/elders (age over 40 years) who lived in Wuhan believed these rumors.

Objective: This study focused on designing a serious game as an experimental program to prevent and control health rumors. The focus of the study was explicitly on the context of the social networking service for midelders/elders.

Methods: This research involved 2 major parts: adopting the Transmission Control Protocol model for games and then, based on the model, designing a game named “Fight With Virus” as an experimental platform and developing a cognitive questionnaire with a 5-point Likert scale. The relevant variables for this experimental study were defined, and 10 hypotheses were proposed and tested with an empirical study. In total, 200 participants were selected for the experiments. By collecting relevant data in the experiments, we conducted statistical observations and comparative analysis to test whether the experimental hypotheses could be proved.

Results: We noted that compared to traditional media, serious games are more capable of inspiring interest in research participants toward their understanding of the knowledge and learning of health commonsense. In judging and recognizing the COVID-19 health rumor, the test group that used game education had a stronger ability regarding identification of the rumor and a higher accuracy rate of identification. Results showed that the more educated midelders/elders are, the more effective they are at using serious games.

Conclusions: Compared to traditional media, serious games can effectively improve midelders’/elders’ cognitive abilities while they face a health rumor. The gameplay effect is related to the individual’s age and educational background, while income and gender have no impact.

(JMIR Serious Games 2024;12:e45546) doi:10.2196/45546

KEYWORDS
serious game; COVID-19; health rumor; game communication; game TCP model; Transmission Control Protocol; gaming; misinformation; disinformation; rumor; health communication; false information; elder; older adult
Introduction

Background

In recent years, the arbitrary spreading of health rumors in mainstream social media on the internet has increasingly gained the attention of the public and raised concerns [1]. For health rumor researchers, a common concern is to propose a feasible and effective prevention and control program for current rampant rumors [2,3]. Furthermore, to prevent and control the spread of such rumors, it is necessary to strengthen the public’s health knowledge to judge and identify the rumors [4]. The concept of a rumor involves a form of statement whose veracity cannot be quickly or ever confirmed. Generally, we have a “dream rumor” and a “bogie rumor,” the former reflecting public desires and wished-for outcomes and the latter hiding some special purpose by somebody, both of them largely occurring during the early COVID-19 epidemic in Wuhan, China [4]. The traditional way of countering rumors often relies on media refutation, which can only be described as a “Band-Aid” solution. To fundamentally prevent and control COVID-19-related rumors and enhance the public’s ability to resist them, we need to find a form of “information vaccine.” Therefore, we chose serious games as a “vaccine” in this context.

The purpose of a serious game is to help people acquire knowledge by playing games. Serious games involve solving problems and studying via careful and thoughtful game ideas [5], while considering characteristics beyond gameplay (eg, purpose and scope [6]). In addition, game elements are used to improve information processing and identify relevant information, which is consistent with the purpose of health rumor prevention research [7]. COVID-19–related rumors are based on the content of serious games to experiment with health rumor prevention, mainly using the control variable method and the analysis-contrast method to apply serious game learning to health rumor prevention research. This paper explores how to help people acquire knowledge of health rumors and health commonsense from the prevention experiment using the relaxed approach of serious games [8].

Previously, serious games have provided a platform for education and business use. For instance, behavioral interventions can be carefully tested and designed to reduce risk-taking behaviors [9], where transmission risks and the usefulness of pandemic-like simulations were demonstrated in the laboratory to be safely and ethically comprehended at the initial state of a health crisis. In addition, other studies prove that serious games are used to accommodate informational and communication complexities in early warning disaster management to simulate and test how public information from social media is used in emergency operation centers to make (protective and communicative) decisions based on levels of trust, usefulness, and completeness [10,11]. Therefore, serious games as an “information vaccine” have certain feasibility, and this paper also explored this issue. Nevertheless, the prevention and control of health rumors have rarely been considered in the context of the social networking service (SNS) for elderly users.

Serious Games

Why are serious games chosen as a solution? Serious games refer to those electronic games whose main content is used for knowledge and skill development, professional training, and spreading culture. They are widely used in many fields. Compared to the limitations and congenital deficiencies of some communication models of traditional media, serious games have become an effective tool to address many social problems, because of their fast speed, wide range, and interactivity [12].

Abt [5] first defined the concept of serious games as follows: “These games have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement.” Later, Sawyer [8], in his white paper titled “Serious Game 2 Initiative,” redefined the concept of a serious game as being an entertainment game with nonentertainment goals. Several variants of the concept have also been proposed. Michael and Chen [13] defined serious games as games that educate, train, and inform. Meanwhile, Zyda [14] defined serious games as a mental contest played with a computer following specific rules. This situation led some analysts to describe serious games as the next wave of technology-mediated learning [15]. Although there is no single definition of the serious game concept, all the proposed definitions convey the same idea: using games to teach or transmit something [16].

Serious games are present in many areas. Westera et al [17] argued that serious games open up many new opportunities for learning complex skills, especially in the education and training domains [12,18-20]. Moreover, Yusoff et al [21] and Crookall [22] argued that good computer games are an excellent example of modern educational theory and that establishing simulation-based serious games as a discipline is a crucial endeavor that could benefit many other related disciplines.

Some early studies were systematically outlined by Connolly et al [23]. For instance, Ziebarth et al [18] and Diehl et al [19] adopted serious games to develop a prototype for the training and education of health students. Some scholars have emphasized the role of serious games in highly specialized skill acquisition (ie, drilling operation [24], mitigation of student dropout [25], improving the command performance of pilots [26]) and education (ie, medical surgery) [27,28], while providing the means to influence cognition and motivational driver [29].

Serious games are also being applied to pass on knowledge or expertise, which can be adopted for various purposes (ie, rehabilitation, psychotherapy, and brain disorders [30-32]). Sometimes, a serious game can also be used to increase risk awareness in the working area of the manufacturing floor [33]. The review by Abd-Alrazaq et al [34] showed that tools such as serious games are usable but are not replaceable options for rehabilitation and clinical intervention where long-term effects are required. Another review by Krath et al [33] revealed that serious games have also incorporated many theoretical foundations relevant to 3 significant landscapes: behavior, learning, and affect-motivation.

In research related to midelders/elders, several studies have demonstrated the potential of serious games to promote physical
activity among older adults. For example, a randomized controlled trial conducted by Fu et al [35] found that a 6-month program of exergaming (exercise using video games) significantly improves the physical function of older adults. Similarly, a study by Jiménez-Pavón et al [36] showed that exergaming increases physical activity and cognitive performance in older adults with mild cognitive impairment.

Serious games have also been used to enhance cognitive training and disease management among older adults. For instance, a study by Anguera et al [37] found that cognitive training through a video game improves cognitive control in older adults. Additionally, a systematic review by Loerzel et al [38] indicated that serious games have the potential to improve self-management and quality of life among older adults with chronic diseases.

In COVID-19–related research, several studies have investigated the potential of gamification and serious games in promoting physical activity during the COVID-19 pandemic. For example, a study by Hall et al [39] proposed a project at a hospital’s senior health center in Canada to discuss how health care can be addressed using serious games among middle-aged and older adults during the pandemic [39]. The study found that the game was effective in increasing physical activity levels and improving self-efficacy. Lau et al [40] demonstrated the potential use of serious game to improve physical activity, cognitive training, and mental health among the aging population during COVID-19 in Hongkong.

Similarly, a study by Suppan et al [41] developed a serious game designed to promote safe behaviors for infection prevention and control (IPC), with a specific focus on COVID-19 among health care workers (HCWs) and other hospital employees. Another study by Ferreira et al [42] explored the potential of gamification in promoting hand hygiene among HCWs during the pandemic. The study found that the game was effective in increasing hand hygiene compliance among the participants [42].

Overall, gamification and serious games have emerged as a promising tool to promote physical activity and health and well-being during the COVID-19 pandemic. These technologies have the potential to support health promotion initiatives and encourage people to adopt healthy behaviors in a fun and engaging way. Therefore, we believe serious games can also solve the issue of COVID-19–related rumors that existed among Chinese midelders/elders.

Health Rumor Analysis

Zhang et al [1] investigated all 453 features of health rumor data collected from a definitive online reference in China. A logistic regression model was adopted to determine the contribution of such features to true and false health rumors. There were measurable differences between true and false health rumors, where the length of a headline or statement and the presence of pictures were negatively correlated with the probability that a rumor was true. Meanwhile, a rumor was more likely to be true if it contained elements such as numbers, source cues, and hyperlinks. They also found that the dread health rumor is more likely to be true than a wishful one. Meanwhile, Chua and Banerjee [4] conducted a study on health rumors from 2015 to 2018. Users’ trust in online health rumors was investigated using 2 factors: length and presence of an image. Additionally, 2 types of rumors were studied: pipe-dream rumor, which offers hope, and bogie rumors, which instill fear. A total of 102 people participated in the experiment, where the finding suggested that pipe-dream rumors are trusted when they are short and do not contain images, while bogie rumors are trusted when they are long and contain images.

Subsequently, Chua and Banerjee [3] investigated the role of epistemic belief in affecting internet users’ decision to share online health rumors. The study focused on the characteristics of rumors—true or false, textual or pictorial, dread or wishful—shaping the decision-making among epistemologically naive and robust users separately. The study showed that epistemologically naive individuals are likelier to share online health rumors than epistemologically robust individuals. In addition, epistemologically robust participants were more likely to share textual rumors than pictorial ones. However, there were no differences between true and false rumors (or between dread and wishful rumors) among either epistemologically naive or robust participants. Meanwhile, Wu [43] modeled factors that predicted fake news sharing during the COVID-19 health crisis. Results showed that informational dependency and social dependency engender both positive and negative cognitive states, namely perceived information timeliness, perceived socialization, and social overload, which then invoke positive and negative affects. Considering that SNS dependency affects information-seeking behavior, it is important for individuals to be exposed to as much accurate information as possible and to build up rational communication against the spread of false rumors.

Ji et al [44] explored factors that influence people’s engagement in scientific rumormongering of genetically modified (GM) food on the Chinese social media platform Sina Weibo at both the group and the individual level. In total, 9070 posts about GM food were obtained from 1 million users. Analysis using logistic regression of the effect of peer influence did not find that users would depend on their friendship network to spread rumors. Instead, results revealed that people with negative attitudes toward GM food and who are social media extroverts (ie, celebrities) are more likely to spread rumors. In contrast, social reputation did not influence the spread of rumors, overwhelming the voices of the scientific community and negatively influencing public attitudes and behaviors.

Meanwhile, Hui et al [45] conducted a study on the spread mechanism of rumors on social network platforms during COVID-19 and considered education as a control measure against the spread of rumors. A novel epidemic-like model was established to characterize the spread of rumors based on 2 dimensions of users (age and time), susceptibility based on education classes, control strategies to effectively restrain rumor propagation, and numerical simulations to verify the main theoretical results. The study concluded that improving education levels and conducting short-term online education are essential strategies for effectively controlling rumor spread. In addition, Pulido et al [46] focused on the social impact of research to identify types of false health information shared on
social media (Reddit, Facebook, and Twitter) using the application of social impact in social media (SISM) methodology. The results indicated that messages focusing on fake health information are primarily aggressive, while those based on the evidence of social impact are respectful and transformative, and deliberation contexts promoted on social media overcome false health information. The findings provide insights into how public health initiatives can support the presence and interactions of evidence as an effective strategy to combat fake news.

A study by Kim and Kim [47] investigated the misinformation belief produced in the context of COVID-19 via 2 main factors: risk perception (psychometric paradigm) and communication. It was found that perceived risk and stigma positively impact belief in fake news, while source credibility and the quantity of information reduce it. Meanwhile, among communication factors, source credibility and the quantity of information reduce belief in fake news, while the credibility of information sources increases it. In addition, Zhao et al [48] used features of online health misinformation that were classified into central level (including topic features) and peripheral level (including linguistic features, sentiment features, and user behavioral features) to propose a health misinformation detection model using the elaboration likelihood model (ELM). Based on a data set collected from a real online health community (because of the lack of a labeled data set), the model correctly detected about 85% of health misinformation. Furthermore, the findings demonstrated the efficacy of behavioral features in health misinformation detection and offered suggestions for misinformation detection by integrating the features of messages and message creators. In COVID-19–related fake news research, Wang and Huang [49] found that although an official denial can initially reduce citizens’ belief in unconfirmed information, later when the denial is revealed to be false, the citizens will have lower levels of belief, not just in the current denial, but also in the government’s future denials of similar rumors. Moreover, the negative lasting effects will carry over to satisfaction with the authorities in the related policy area.

COVID-19 Background

This paper was initially written in 2020, and the experiment was conducted in the period from February to March 2020. Therefore, many things changed from then up to the Omicron strain of COVID-19. As such, we acknowledge that this paper has time constraints; however, the research still provides some valuable inspiration and conclusions on game studies, media development, and health care. Since the COVID-19 pandemic broke out in December 2019, the related health rumors also began to wreak havoc on the internet.

Rumor prevention is difficult in the case of rumors that rely on propaganda, and the educational means of traditional media are ineffective due to the lack of interaction and the complexity of information. On the internet, especially the midelders/elders were in a state of panic and information-blind obedience [50].

In China, an SNS group existed, in addition to many WeChat groups, similar to Discord and Facebook in the West. Therefore, we could easily find a target sample for our newly established experimental community, where any questions could be communicated at any time. Our experiment was conducted in early 2020, and some people could answer the questionnaire face to face, while others could not because of the lockdown. Therefore, some respondents were sent offline paper questionnaires, and we also requested them to fill in the online questionnaire. The Chinese midelders/elders were comfortable playing the game on their cell phones, so they easily believed the health rumor that the information communication channel is too fast. Some of them whom we could not meet face to face were contacted over a video call, and we confirmed their age and other personal information clearly to ensure accuracy in the experiment.

The original survey, questionnaire, and serious game are in Chinese, convenient for our non–English-speaking respondents, and all the concepts in this paper are the translated version. This means we just translated the statistical data and labels; during the experiment, there was no translation, and we followed the same steps for all the scales.

**Elderly WeChat Users in China: Original Survey**

According to our data collected in the original survey, the contemporary middle-aged and older adults, especially those aged 40-60 years, have a high frequency of use of WeChat; the number of elderly WeChat users with frequent use accounts for 66.09% of the total. According to interviews at different levels of the questionnaire survey process, middle-aged and elderly users of WeChat are aged from 50 to 65 years. They are also familiar with using the WeChat “circle of friends” function and other social media platforms (eg, TikTok). They often record their daily lives and travel through videos and pictures. Generally, this user group is also active in online social group chats, and their frequency of using online social media is no less than that of some young user groups. For example, 48.85% of middle-aged and elderly WeChat users said they occasionally read health information on WeChat, and only 16.67% said they had never received health information forwarded by relatives and friends (Figure 1).

The survey on the acquisition and dissemination of health information by elderly WeChat users was the focus of this study. Most people do not have the habit of reading health information regularly. It can be seen from the data that this depends to some extent on the frequency of obtaining information. People read health information only when it is forwarded to them by relatives and friends or when relevant health public accounts push this information or when it is in the form of characteristic health information news, as shown in Figure 2. Regarding access to health information, 62.07% of the respondents received health information from their WeChat friends. In addition, 83.33% of the respondents had the experience of forwarding health information to their children or parents, and 57.47% of those forwarded health information to their WeChat friends. Most respondents felt that the original intention of forwarding health information was to help others with a positive attitude.
However, many midelders/elders received health information without any judgment and recognition and then spread the information with a “good intention” motive, which is also why the health rumor issue is rampant. The data also show that the failure to recognize and identify health rumors is more likely to be the reason than the motive for spreading them. In addition, according to the questionnaire, 86.78% of elderly WeChat users trusted health information forwarded by relatives and friends and 45.4% considered it very trustworthy. The trustworthiness of health information forwarded by colleagues was 82.18%. These data show that WeChat has become a hotbed for health rumors among the midelders/elders.

Therefore, this paper used a serious game as a tool to test the effect of game media on the prevention of health rumors. Compared with other media, the serious game had a special communication model and effect that could improve this situation (see the Results section for more details). Therefore, using the COVID-19 pneumonia rumor was suitable as the target and content of the serious game, involving not only the elderly closely related to COVID-19 pneumonia but also COVID-19 rumor communication relying on WeChat. Finally, the number of health rumors that emerged during the COVID-19 epidemic was enormous, and enough rumor cases could be collected for experimentation.

**Methods**

**The Transmission Control Protocol Model of the Game**

There are many theoretical models concerning the communication effect of games as media [51]. The computer networking concept was adopted as the inspiration for this research based on the idea of engineering. Two main protocols exist in network communication: Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) [51]. TCP originated in the initial network implementation, complementing Internet Protocol (IP). TCP provides reliable, ordered, and error-checked delivery of a stream of bytes between applications running on hosts communicating via an IP network. Major internet applications, such as the World Wide Web, email, remote administration, and file transfer, rely on TCP because of the 3-way handshake mechanism (Figure 3) [52].

Having introduced the logical mechanism of UDP and TCP from the technical perspective of communication, we can see that all media communication models are suited to TCP and UDP (2 computer network theories). UDP uses a simple connectionless communication model, just from the information source to the information sink. For example, the newspaper provides information to readers without any interaction (request and response). However, for all current media, only games
match the TCP model (Figure 4). In traditional media, no matter
the newspaper, broadcaster, or television program, the audience
only receives information; the UDP model does not have a
feedback process, the timeliness is good, but transmission is
unstable. As a result, users can refuse to accept information or
hardly notice useful content. Therefore, serious games can help
society to address health rumor issues. In this paper, we proved
the effect on the health communication area [53].

Figure 3. Three-way handshaking in TCP. ACK: acknowledge; RTT: round-trip time; SYN: synchronize; TCP: Transmission Control Protocol.

Figure 4. Serious game of TCP. RTT: round-trip time; TCP: Transmission Control Protocol.
Study Design
This study investigated the prevention and control of health rumors in WeChat, as most elderly WeChat users are concerned about health information and are negatively affected by health rumors. Here, the term “elderly” in our paper is a macroscopic definition: it is not only a physiological age classification but also a description of the psychology or state. In China, people who believe a health rumor via the SNS in the age range of 40-60 years (midelders/elders) were considered. We recruited 200 midelders/elders in Tongren City, Guizhou Province, China, which did not have a serious spread of COVID-19 in early 2020. The participants gathered to dance and train in the city plaza, and then, we requested them to attend our game experiment.

The experimental program was constructed in 2 parts. The first part was developing a serious game based on the content of health rumors and health commonsense; we named it “Fight With Virus.” The purpose was to apply this in a health rumor prevention experiment. The second part was developing a cognitive questionnaire with the theme of COVID-19 health rumor, with 5-point Likert scale, which aimed to compare and analyze the prevention effects of traditional and serious game learning models on health rumors. Baishya and Samalia [54] extended the unified theory of acceptance and use of technology (UTAUT) into UTAUT2, incorporating 3 constructs into the original UTAUT: hedonic motivation, price value, and habit. Individual differences (ie, age, gender, and experience) were hypothesized to moderate the effects of these constructs on behavioral intention and technology use, thus affecting their learning of new technologies. Therefore, according to several past studies based on the UTAUT2 model [55,56], this study adopted the UTAUT2 model to analyze the effect of the serious game. We modified and added variables, which were analyzed using IBM SPSS Amos and IBM SPSS Statistics on factors influencing health information use and dissemination. On this basis, a suitable serious game experiment scheme was built.

The experiment was conducted in 4 steps. In steps 1 and 2, we selected the target participants (midelders/elders), while in steps 3 and 4, we designed the game for the experiment.

Step 1
The construction of the experimental program based on serious games and experimental research needed to be based on a full understanding of the use and dissemination of health information by the research participants. We analyzed the health information needs of the research participants, the frequency and channels of use and the dissemination of health information, and their ability to identify and judge health rumors.

Step 2
To investigate the phenomenon of the dissemination of health information in WeChat’s midelder/elder user groups, we used a questionnaire designed in 3 parts: The first part involved a survey to collect personal information, such as gender, age, place of residence, income level, and education. The second part was a survey on the habit of using WeChat. The third part mainly involved the frequency, channel, and motivation of users to obtain and forward health information.

At the same time, 30 health rumor judgment questions were attached to this survey questionnaire, and respondents were asked to judge whether they were correct or incorrect. Through the correct rate of health rumor judgment, we determined the trust level and ability of the respondents to identify health rumors. We also popularized the 30 relevant health rumors, with the hope to popularize the degree of health rumor knowledge and also to strengthen the respondents’ ability to recognize information. The questionnaire is shown in Multimedia Appendix 1.

Based on the cognitive ability determined through the questionnaire, 200 participants were selected and asked for their willingness to play the serious game.

Step 3
Based on the use and dissemination of health information by the research participants, the theme of the health rumor learning content was selected and a serious game experimental scheme suitable for this group was constructed through the design and production of serious game content. Considering the experimental length of the serious game and the understanding and acceptance level of the participants, the video game mode of a multiline plot was not applicable for our research, so a single-line plot and scenario was used in the design of the game.

The learning content of the serious game is mainly based on the theme of “a personal day,” and the content of the game plot is a person’s life from morning to afternoon, in the form of a single storyline. An explanation is provided at the beginning of the game to accurately communicate the theme, rules, and intent of the game to the players. In the learning content of the game, information such as health rumors and general knowledge about COVID-19 was selected, as shown in Table 1, and based on the selected content, failure/passing conditions were set for the game, which involved “risk of infection” and “psychological stress.” Different scenarios are set up in a day’s life, and questions are set up to interact with the game players to promote and increase the knowledge of COVID-19-related rumors in this interactive learning serious game. The main line design is shown in Figure 5, and the game logic is shown in Figure 6.

In this study, to achieve the effect of the serious game and the purpose of health rumor prevention, a feedback link of the serious game–based health rumor control prevention experiment was important. The feedback link was mainly achieved by setting up a feedback mechanism, which reflected the understanding of the research participants (players) of the game content (COVID-19 health rumor); by setting up the feedback mechanism, interactivity with the research participants could also be strengthened. At the same time, the feedback data were used to reflect the learning effect of the serious game.

The feedback mechanism of the serious game–based health rumor prevention experiment was implemented in the following 3 parts.

- The first part was to communicate the theme, rules, and intention of the game to the research participants by means of game instructions at the beginning of the game. This is an important step to quickly integrate the player into the...
learning process of a serious game and to let the player know what they will do next in the game.

• The second part was realized in the textual feedback of the gameplay process, where the player was provided with choices through interactive video scenarios, and instant feedback was provided. Instant feedback is an important part of the overall feedback process, which needs to be clearly communicated to the player. It is necessary to clearly communicate to the player whether their choices are correct and to strengthen the knowledge of health rumors and general health. The textual feedback content of the game process is shown in Table 2.

• The last part was to provide feedback after the player passed or failed in the game. At the end of the serious game, based on the player’s overall understanding of COVID-19 health rumors and health knowledge, the feedback can strengthen the player’s knowledge of health rumors.

The serious game created in this study used a COVID-19 health rumor as the learning content (see Figures 7 and 8). To achieve the purpose of preventing and controlling health rumors, a textual feedback mechanism was designed, involving 4 infection risks and 7 psychological stress settings. These were assigned to game failure or passing conditions, as shown in Figures 9 and 10. The game data reflected the performance of the research participant (player) in the game, with the settings shown in Table 3 to cater to the experiment’s needs. The specific game data value settings and game passing/failure conditions are shown in Figure 11. Specifically, the story is as follows: The protagonist, a young person, suddenly finds themselves caught up in the COVID-19 pandemic in early 2020 in their city. Various pieces of information related to COVID-19 start to emerge around the protagonist, causing a massive explosion of fear and panic, particularly among many elderly people who turn to social media for information. They begin to demand that the protagonist follow their advice on preventing the pandemic. The goal of the game is to distinguish between real health knowledge and rumors throughout the daily life story, to use accurate knowledge to save the elderly citizens who are in a state of panic, and to slow down the spread of the virus. In the end, the game outcome is judged based on the actions and choices made by the protagonist.

Table 1. Selection of health rumors/health facts for the serious game content.

<table>
<thead>
<tr>
<th>Time of day and COVID-19 health rumor/health fact</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Morning</strong></td>
<td></td>
</tr>
<tr>
<td>R1: Drinking plenty of boiled water at 60°C can prevent COVID-19.</td>
<td>Infection risk</td>
</tr>
<tr>
<td>R2: Domesticated dogs and cats can also spread COVID-19.</td>
<td>Psychological stress</td>
</tr>
<tr>
<td>R3: Putting used masks in a sterilizer can continue to provide protection and use.</td>
<td>Infection risk</td>
</tr>
<tr>
<td>The correct way to wear a mask (not an option).</td>
<td>General health knowledge</td>
</tr>
<tr>
<td>R4: Going out with ginger slices in the mouth can prevent COVID-19.</td>
<td>Psychological stress</td>
</tr>
<tr>
<td>R5: The government will use military aircraft to spread disinfectants in the sky.</td>
<td>Psychological stress</td>
</tr>
<tr>
<td><strong>Noon</strong></td>
<td></td>
</tr>
<tr>
<td>R6: You should keep more than 1 m distance from strangers when you go out in times of an epidemic.</td>
<td>Infection risk</td>
</tr>
<tr>
<td>R7: Eye-to-eye contact may transmit COVID-19.</td>
<td>Psychological stress</td>
</tr>
<tr>
<td>R8: Shuanghuanglian Oral Liquid can effectively inhibit the COVID-19 virus.</td>
<td>Psychological stress</td>
</tr>
<tr>
<td><strong>Afternoon</strong></td>
<td></td>
</tr>
<tr>
<td>R9: Disinfection is required for items after returning home from outside.</td>
<td>Infection risk</td>
</tr>
<tr>
<td>R10: High temperature can kill the virus, so hot blow-drying and hot water bathing can inhibit it.</td>
<td>Psychological stress</td>
</tr>
<tr>
<td>R11: Do not eat fish; pickled fish made from grass carp can transmit COVID-19.</td>
<td>Psychological stress</td>
</tr>
</tbody>
</table>

aR: rumor.
Figure 5. The game process. R: rumor.
Table 2. Text feedback during gameplay.

<table>
<thead>
<tr>
<th>Rumor number</th>
<th>Textual feedback (explanation and education) of scenario options during gameplay</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Drinking water does not help, and scalding the mucous membrane of the mouth with hot water can increase the risk of infection.</td>
</tr>
<tr>
<td>R2</td>
<td>There is no evidence that COVID-19 can be transmitted to domesticated dogs and cats.</td>
</tr>
<tr>
<td>R3</td>
<td>Masks that have been used many times do not work to isolate droplets.</td>
</tr>
<tr>
<td>R4</td>
<td>Graphic feedback: follow the 3 steps (regulations) to wear the mask correctly.</td>
</tr>
<tr>
<td>R5</td>
<td>Ginger does not work to prevent the COVID-19.</td>
</tr>
<tr>
<td>R6</td>
<td>This is a rumor. There are no military aircraft to spread disinfectants in the sky; in addition, the local government has no right to do that. (This raises the players' sense of alertness and achieves the purpose of public education.)</td>
</tr>
<tr>
<td>R7</td>
<td>COVID-19 is spread via droplets and contact, and close contact increases the risk of infection.</td>
</tr>
<tr>
<td>R8</td>
<td>This is a rumor. The virus is transmitted through bodily fluids, droplets, and aerosols, not through the eyes. (This re-explains the mode of transmission of the COVID-19 virus.)</td>
</tr>
<tr>
<td>R9</td>
<td>Clearly inform that this is not yet clear information and should not be followed blindly.</td>
</tr>
<tr>
<td>R10</td>
<td>This is true. You should do that. (This provides possible contact transmission and health information on sterilization.)</td>
</tr>
<tr>
<td>R11</td>
<td>COVID-19 only infects mammals. Fish do not transmit COVID-19.</td>
</tr>
</tbody>
</table>

*aR: rumor.  
*bNot applicable.*
**Figure 7.** The experimental serious game’s gameplay content 1 (Chinese version). Question: “Hi boy, do you know where one can buy Shuanghuanglian Oral Liquid (a Chinese medicine)? I hear it is useful for COVID-19 treatment!” Answer options: (A) “Really? I also want to buy some.” (B) “We do not know the drug’s action, so do not drink it by yourself!”.

**Figure 8.** The experimental serious game’s gameplay content 2 (Chinese version). Question: “Please come back to home soon; the government will use military aircraft to spray disinfectants!” Answer options: (A) “Really? I’m leaving right now.” (B) “Fake news, Mom!”.

**Figure 9.** The experimental serious game: game over (Chinese version). Meaning: “The psychological pressure is 10, the risk of infection is 40, you are in a high-risk situation!”.
**Figure 10.** The preventive knowledge statement after the game, explaining how to wear a mask in 3 steps.

**Table 3.** Game data and game rating settings.

<table>
<thead>
<tr>
<th>Game score</th>
<th>Error choice</th>
<th>Game failure</th>
<th>Game round</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>≤3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>4-6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>7-9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>≥10</td>
<td>≥3</td>
<td>≥4</td>
</tr>
</tbody>
</table>

**Figure 11.** Game-related data values and failure/passing condition-setting thinking diagram.
Step 4

Finally, the data and feedback of the participants were obtained, and the effect of the serious game on health rumor prevention was analyzed through the data and feedback to determine whether serious games are useful to prevent health rumors.

Data Collected

This study compared and analyzed the differences between acquiring and understanding health rumor information through the learning modes of serious games and traditional media. A total of 100 people were selected to participate in the serious game experiment (G1 group), while 100 people who did not participate in the serious game experiment only studied by traditional media (G2 group). To ensure the objectivity of the controlled experiment, the educational background of the 200 participants was investigated before the formal study while keeping the 2 groups as similar as possible in terms of gender and age, as Table 4 shows. Next, we sent the testing questionnaire related to health commonsense and health rumors to G1 and G2.

Table 4. Demographic information of groups G1 and G2.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>G1 (n=100), n (%)</th>
<th>G2 (n=100), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>42 (42)</td>
<td>40 (40)</td>
</tr>
<tr>
<td>Female</td>
<td>58 (58)</td>
<td>60 (60)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-45</td>
<td>24 (24)</td>
<td>27 (27)</td>
</tr>
<tr>
<td>46-50</td>
<td>30 (30)</td>
<td>25 (25)</td>
</tr>
<tr>
<td>51-55</td>
<td>19 (19)</td>
<td>17 (17)</td>
</tr>
<tr>
<td>56-60</td>
<td>18 (18)</td>
<td>21 (21)</td>
</tr>
<tr>
<td>≥61</td>
<td>9 (9)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>Educational background</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>15 (15)</td>
<td>15 (15)</td>
</tr>
<tr>
<td>High school/technical secondary school</td>
<td>50 (50)</td>
<td>50 (50)</td>
</tr>
<tr>
<td>Junior college</td>
<td>20 (20)</td>
<td>20 (20)</td>
</tr>
<tr>
<td>Bachelor’s degree and higher</td>
<td>15 (15)</td>
<td>15 (15)</td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>29 (29)</td>
<td>24 (24)</td>
</tr>
<tr>
<td>Average</td>
<td>48 (48)</td>
<td>55 (55)</td>
</tr>
<tr>
<td>High</td>
<td>23 (23)</td>
<td>21 (21)</td>
</tr>
</tbody>
</table>

Variables and Hypotheses

Based on the framework of serious games, a questionnaire was designed that included the following variable definitions:

- Independent variables: The learning mode refers to the way that information is obtained and knowledge learned; the variables were M1 (learning through serious games) and M2 (learning through traditional and new media). For gender, age, income, and education, in the preparation stage of the research, education was used as the main grouping basis, and the age distribution and gender ratio of the 2 groups of experimental objects were kept consistent.

- Intermediary variables: Game data for serious game experiments (A/B/C/D) refer to the player’s performance data during the game, including the number of selection errors, the number of game failures, and the number of game passings. Personal performance was divided into 4 mediating variables: excellent (A), good (B), medium (C), and poor (D).

- Dependent variables: The variables were cognitive questionnaire (1) overall correct response rate of judgment and recognition of the COVID-19 health rumor (X1 for G1, Y1 for G2 [G2 did not participate in the serious game experiment]), (2) correct rate of judgment and recognition the COVID-19 health rumor part 1 (X2 for G1 [COVID-19 health rumor not included in the serious game experiment], Y2 for G2), and (3) correct rate of judgment and identification of the COVID-19 health rumor part 2 (X3 for G1, Y3 for G2 [COVID-19 pneumonia rumor included in the traditional media experiment for G2]).

- Intervening variables: The comprehension, cognitive level, learning ability, learning interest, and information attention of the G1 group affected the outcome of the dependent variables to a certain extent. The specific influencing relationship between various variables is shown in Figure 12.
The following 10 hypotheses were proposed in this experimental study:

- **Hypothesis 1 (H1)**: Serious game experiments can help research participants acquire and understand health rumor knowledge and health commonsense.
- **H2**: The serious game learning mode is more capable of inspiring the interest of research participants in their understanding of the knowledge acquired and their learning of health commonsense compared to the traditional learning mode.
- **H3**: The serious game learning mode is more impactful than the traditional learning mode.
- **H4**: In judging and recognizing the COVID-19 health rumor, G1 has a stronger judgment ability than G2 and a higher accuracy in identifying the rumor in the serious game experiment.
- **H5**: In judging the COVID-19 health rumor, for the rumor not included in the serious game experiment, without the influence of M1, the ability of G1 and G2 is not much different, and the accuracy rate of identifying the COVID-19 health rumor is roughly the same for both groups. The manifestation in the variable is $X_2 = Y_2$.
- **H6**: In judging and recognizing the COVID-19 health rumor, G1 has an overall stronger judgment ability than G2 and a higher accuracy rate of identifying the COVID-19 health rumor. The specific manifestation in the variable is $X_3 > Y_3$.
- **H7**: Gender affects G1’s and G2’s judgment and recognition of the COVID-19 health rumor.
- **H8**: Age affects G1’s and G2’s judgment and recognition of the COVID-19 health rumor.
- **H9**: Income affects G1’s and G2’s judgment and recognition of the COVID-19 health rumor.
- **H10**: Academic qualifications affect G1’s and G2’s judgment and recognition of the COVID-19 health rumor.

**Ethical Considerations**

According to the guidelines of the People’s Republic of China [57], this study met the conditions for exemption from ethical review.
Results

Analysis of Data Collected

According to the collected game data, 36% (72/200) of players received A, 41% (82/200) received B, 18% (36/200) received C, and the remaining 5% (10/200) received D. The accuracy rate of the judgment and recognition of the COVID-19 health rumor and health commonsense in the cognitive questionnaire of G1 and G2 groups are tabulated in Table 5. The overall accuracy was 84% for G1 and 78% for G2, with an average of 81%. The relationship of the parameters X1-X3 (G1) and Y1-Y3 (G2) are shown in Figures 13 and 14, respectively.

Table 5. Cognitive questionnaire data (judgment and recognition of rumor knowledge and health commonsense).a

<table>
<thead>
<tr>
<th>Question number</th>
<th>Accuracy G1 (%)</th>
<th>Accuracy G2 (%)</th>
<th>Average accuracy (%), (G1+G2)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72</td>
<td>76</td>
<td>74</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>98</td>
<td>96</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>79</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>77</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td>5</td>
<td>64</td>
<td>63</td>
<td>64</td>
</tr>
<tr>
<td>6</td>
<td>87</td>
<td>91</td>
<td>89</td>
</tr>
<tr>
<td>7</td>
<td>77</td>
<td>85</td>
<td>81</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>70</td>
<td>71</td>
</tr>
<tr>
<td>9</td>
<td>67</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>80</td>
<td>74</td>
<td>77</td>
</tr>
<tr>
<td>11</td>
<td>67</td>
<td>64</td>
<td>66</td>
</tr>
<tr>
<td>12</td>
<td>80</td>
<td>88</td>
<td>84</td>
</tr>
<tr>
<td>13</td>
<td>88</td>
<td>96</td>
<td>92</td>
</tr>
<tr>
<td>14</td>
<td>75</td>
<td>71</td>
<td>73</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>16</td>
<td>91</td>
<td>95</td>
<td>93</td>
</tr>
<tr>
<td>17</td>
<td>68</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>18</td>
<td>80</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>19</td>
<td>58</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>20</td>
<td>97</td>
<td>71</td>
<td>84</td>
</tr>
<tr>
<td>21</td>
<td>99</td>
<td>61</td>
<td>80</td>
</tr>
<tr>
<td>22</td>
<td>98</td>
<td>65</td>
<td>82</td>
</tr>
<tr>
<td>23</td>
<td>99</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td>24</td>
<td>91</td>
<td>81</td>
<td>86</td>
</tr>
<tr>
<td>25</td>
<td>100</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td>26</td>
<td>100</td>
<td>89</td>
<td>95</td>
</tr>
<tr>
<td>27</td>
<td>90</td>
<td>76</td>
<td>83</td>
</tr>
<tr>
<td>28</td>
<td>92</td>
<td>85</td>
<td>89</td>
</tr>
<tr>
<td>29</td>
<td>92</td>
<td>67</td>
<td>80</td>
</tr>
<tr>
<td>30</td>
<td>99</td>
<td>95</td>
<td>97</td>
</tr>
</tbody>
</table>

*aAll the correct rate values are kept to integer bits.
According to the data, the overall average correct rate of judgment and recognition of G1 and G2 was 84% and 78%, respectively. In addition, the related data collected showed that 76% of the participants believed that the serious game learning mode is more interesting than the traditional learning mode. This finding shows that G1 has good interest in games. Furthermore, 60% of the participants thought that the game learning mode is more helpful than the traditional learning mode, and 65% thought that the serious game learning mode makes a more profound impression. Therefore, from the perspective of the selection of participants, the serious game learning mode is more interesting, helpful, and impactful than the traditional learning mode. Therefore, H2 and H3 hold.

**Analysis of Dependent Variables (X, Y) and Intermediary Variables (A, B, C, D)**

According to the data of the dependent variables in Table 6, we could not directly prove that the impact of M1 on X1 was greater than that of M2 on Y1. At the same time, when the knowledge of health rumors and health commonsense was not included in the game content, the correct rate of judgment and recognition of G1 and G2 was almost the same, and even G1 had a relatively lower rate. However, X3 was 96%, which is much higher than Y3 (79%), and X3 exceeded X2 by up to 23 percentage points. This condition implies that the impact of M1 on X1 is greater than the impact of M2 on Y1. It not only shows that G1 had relatively strong learning ability but also that after the serious game learning model experiment, there was a significant positive effect on the accuracy rate of the judgment and recognition accuracy of the COVID-19 health rumor and health commonsense. At the same time, X1 > Y1, X2 ≈ Y2, and X3 > Y3. Therefore, H4, H5, and H6 are established.

The intermediary variables A, B, C, and D were sequentially observed, corresponding to the dependent variables X1, X2, and X3. It can be clearly seen in Table 7 that A-X1 > B-X1 > C-X1 > D-X1, A-X2 > B-X2 > C-X2 > D-X2, and A-X3 > B-X3 > C-X3 > D-X3. The dependent variables corresponding to the intermediary variables showed a decreasing trend from A to D, indicating that M1 affected X1 and X3 through A, B, C, and D and the degree of influence was in the order of A > B > C > D. These data showed that the higher the average game score, the higher the correct rate of recognition and judgment. Therefore, combined with the previous analysis, H1 holds.
Table 6. Dependent variables X and Y.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>84</td>
</tr>
<tr>
<td>X2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>76</td>
</tr>
<tr>
<td>X3&lt;sup&gt;c&lt;/sup&gt;</td>
<td>96</td>
</tr>
<tr>
<td>Y1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>78</td>
</tr>
<tr>
<td>Y2&lt;sup&gt;e&lt;/sup&gt;</td>
<td>78</td>
</tr>
<tr>
<td>Y3&lt;sup&gt;f&lt;/sup&gt;</td>
<td>79</td>
</tr>
</tbody>
</table>

<sup>a</sup>X1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G1.
<sup>b</sup>X2: cognitive questionnaire correct rate of judgment and recognition part 1 for G1.
<sup>c</sup>X3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G1.
<sup>d</sup>Y1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G2.
<sup>e</sup>Y2: cognitive questionnaire correct rate of judgment and recognition part 1 for G2.
<sup>f</sup>Y3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G2.

Table 7. Intermediary variables (A, B, C, D) and dependent variables (X1-X3).

<table>
<thead>
<tr>
<th>Intermediary variable</th>
<th>Dependent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X1&lt;sup&gt;a&lt;/sup&gt; (%)</td>
</tr>
<tr>
<td>A</td>
<td>87</td>
</tr>
<tr>
<td>B</td>
<td>83</td>
</tr>
<tr>
<td>C</td>
<td>80</td>
</tr>
<tr>
<td>D</td>
<td>72</td>
</tr>
</tbody>
</table>

<sup>a</sup>X1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G1.
<sup>b</sup>X2: cognitive questionnaire correct rate of judgment and recognition part 1 for G1.
<sup>c</sup>X3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G1.

Analysis of Independent Variables

**Gender**

We grouped participants, ensuring the educational composition of the 2 groups was as consistent as possible. By observing and comparing the G1 independent variable (gender) and its corresponding intermediary variables (Table 8), we found that the game score reached A, where the male participants were better than the female participants but changed from B to D when the female participants were better than the male participants. This condition was especially true when the game score reached B, where the female participants were much better than the male participants. This situation may also be influenced by the unequal relationship of the overall gender. There was no gender difference in the numbers from game rating A to D.

Subsequently, by observing and comparing the G1 and G2 independent variable gender and the corresponding dependent variables (Table 9), we observed that regarding the dependent variables X1, X2, and X3, corresponding to the independent variable gender (female, male), the comparisons were female<male, female<male, and female>male respectively. Regarding Y1, Y2, and Y3, corresponding to gender, the comparisons were female<male, female>male, and female<male respectively. As such, there was no gender difference. Therefore, H7 does not hold.

Table 8. G1 gender and corresponding intermediary variables.

<table>
<thead>
<tr>
<th>Intermediary variable</th>
<th>Gender</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>A</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>27</td>
<td>14</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 9. G1 and G2 gender and corresponding dependent variables.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Gender</th>
<th></th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female (%)</td>
<td>Male (%)</td>
<td></td>
</tr>
<tr>
<td>X1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>83</td>
<td>84</td>
<td>Female&lt;male</td>
</tr>
<tr>
<td>X2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>75</td>
<td>77</td>
<td>Female&lt;male</td>
</tr>
<tr>
<td>X3&lt;sup&gt;c&lt;/sup&gt;</td>
<td>97</td>
<td>95</td>
<td>Female&gt;male</td>
</tr>
<tr>
<td>Y1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>76</td>
<td>81</td>
<td>Female&lt;male</td>
</tr>
<tr>
<td>Y2&lt;sup&gt;e&lt;/sup&gt;</td>
<td>79</td>
<td>77</td>
<td>Female&lt;male</td>
</tr>
<tr>
<td>Y3&lt;sup&gt;f&lt;/sup&gt;</td>
<td>71</td>
<td>88</td>
<td>Female&lt;male</td>
</tr>
</tbody>
</table>

<sup>a</sup>X1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G1.
<sup>b</sup>X2: cognitive questionnaire correct rate of judgment and recognition part 1 for G1.
<sup>c</sup>X3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G1.
<sup>d</sup>Y1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G2.
<sup>e</sup>Y2: cognitive questionnaire correct rate of judgment and recognition part 1 for G2.
<sup>f</sup>Y3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G2.

Age

There were apparent differences in the age ranges between the 2 groups, as shown in Table 4. Therefore, random sampling in G1 and G2 was conducted, and 25 participants under 51 and 56 years old each were selected, with 50 participants in each group for comparative observation and analysis of the corresponding variable data.

First, by observing and comparing the high-age and low-age groups’ independent and variable age groups and their corresponding intermediary variables (Table 10), we found that the number of people who achieved A and B game scores were all of low age. As a result, the number of people in the low-age group was greater than the number of people in the high-age group; among those with game scores C and D, the number of people in the high-age group was greater than the number of people in the low-age groups, indicating that to a certain extent, the independent variable age positively affects the intermediary variables A, B, C, and D. Second, by observing and comparing the high- and low-age groups of the G1 and G2 independent variable age with corresponding dependent variables (Table 11), the values of independent variables X1, X2, and X3 could be determined. The values of the low-age group were greater than those of the high-age group; the independent variables Y1, Y2, and Y3 also exhibited this behavior. Therefore, H8 holds.

Table 10. G1 and G2 targets of different ages.

<table>
<thead>
<tr>
<th>Intermediary variable</th>
<th>Age</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low (n=25), n (%)</td>
<td>High (n=25), n (%)</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>8 (32)</td>
<td>13 (52)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>6 (24)</td>
<td>8 (32)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>8 (32)</td>
<td>3 (12)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>3 (12)</td>
<td>1 (4)</td>
<td></td>
</tr>
</tbody>
</table>
### Table 11. G1 and G2 age groups and corresponding dependent variables.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Age</th>
<th>Low (%)</th>
<th>High (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1\textsuperscript{a}</td>
<td>76</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>X2\textsuperscript{b}</td>
<td>69</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>X3\textsuperscript{c}</td>
<td>88</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Y1\textsuperscript{d}</td>
<td>68</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Y2\textsuperscript{e}</td>
<td>66</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Y3\textsuperscript{f}</td>
<td>72</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}X1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G1.
\textsuperscript{b}X2: cognitive questionnaire correct rate of judgment and recognition part 1 for G1.
\textsuperscript{c}X3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G1.
\textsuperscript{d}Y1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G2.
\textsuperscript{e}Y2: cognitive questionnaire correct rate of judgment and recognition part 1 for G2.
\textsuperscript{f}Y3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G2.

### Income

The relevant data collected are shown in Table 4. Nearly half of the participants in G1 and G2 believed that their income level was average, and the number of people who believed that their income was high or low was relatively small. By observing and comparing the high- and low-income subgroups in G1 and G2 with corresponding dependent variables, we found that the dependent variables corresponding to the 2 independent variable subgroups were not identical, as shown in Table 12. Therefore, H9 does not hold.

### Table 12. High- and low-income groups and corresponding dependent variables of G1 and G2.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Income</th>
<th>Low (%)</th>
<th>High (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1\textsuperscript{a}</td>
<td>High (%)</td>
<td>84</td>
<td>84</td>
</tr>
<tr>
<td>X2\textsuperscript{b}</td>
<td>High (%)</td>
<td>82</td>
<td>78</td>
</tr>
<tr>
<td>X3\textsuperscript{c}</td>
<td>High (%)</td>
<td>88</td>
<td>94</td>
</tr>
<tr>
<td>Y1\textsuperscript{d}</td>
<td>High (%)</td>
<td>73</td>
<td>74</td>
</tr>
<tr>
<td>Y2\textsuperscript{e}</td>
<td>High (%)</td>
<td>69</td>
<td>72</td>
</tr>
<tr>
<td>Y3\textsuperscript{f}</td>
<td>High (%)</td>
<td>80</td>
<td>78</td>
</tr>
</tbody>
</table>

\textsuperscript{a}X1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G1.
\textsuperscript{b}X2: cognitive questionnaire correct rate of judgment and recognition part 1 for G1.
\textsuperscript{c}X3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G1.
\textsuperscript{d}Y1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G2.
\textsuperscript{e}Y2: cognitive questionnaire correct rate of judgment and recognition part 1 for G2.
\textsuperscript{f}Y3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G2.

### Education

Based on the cognitive questionnaire, the independent variable education was divided into 4 segments, as shown in Table 4. First, 4 groups of the G1 independent variable education and corresponding intermediary variables were compared (Table 13). Through comparison and observation, we found that the higher the education level, the better the performance in the game, which demonstrates that the dependent variable education positively affects the intermediary variables. Second, the grouping and corresponding dependent variables of G1 and G2 based on academic qualifications are shown in Table 14. We found that education has a positive effect on the corresponding dependent variables. Therefore, H10 holds.

Finally, a thorough investigation of the selection tendency between the learning modes of serious games (M1) and traditional learning (M2) was conducted on G1 involving a comparison experiment of interest, help, and impression (Figure 15). The data showed that 76% (152/200) of the participants thought the serious game learning mode was more interesting than the traditional learning mode, indicating that G1 had a reasonable learning interest in games. Furthermore, 60%
(120/200) of the participants thought that the serious game learning mode was more helpful than the traditional learning mode, and 65% (130/200) thought that the serious game learning mode was more impressive than the traditional learning mode. Therefore, from the perspective of the selection tendency of the participants, the serious game learning mode is more interesting, helpful, and impressive than the traditional learning mode, which strengthens the hypotheses.

**Table 13.** Proportion of game scores in different education segments.

<table>
<thead>
<tr>
<th>Segment</th>
<th>A (%)</th>
<th>B (%)</th>
<th>C (%)</th>
<th>D (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>20</td>
<td>47</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>High school</td>
<td>28</td>
<td>44</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>College degree</td>
<td>45</td>
<td>40</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Bachelor’s degree and higher</td>
<td>59</td>
<td>23</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 14.** Education segments and corresponding dependent variables of G1 and G2.

<table>
<thead>
<tr>
<th>Segment</th>
<th>X1&lt;sup&gt;a&lt;/sup&gt; (%)</th>
<th>X2&lt;sup&gt;b&lt;/sup&gt; (%)</th>
<th>X3&lt;sup&gt;c&lt;/sup&gt; (%)</th>
<th>Y1&lt;sup&gt;d&lt;/sup&gt; (%)</th>
<th>Y2&lt;sup&gt;e&lt;/sup&gt; (%)</th>
<th>Y3&lt;sup&gt;f&lt;/sup&gt; (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>83</td>
<td>72</td>
<td>94</td>
<td>72</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td>High school</td>
<td>82</td>
<td>74</td>
<td>95</td>
<td>77</td>
<td>77</td>
<td>78</td>
</tr>
<tr>
<td>College degree</td>
<td>86</td>
<td>79</td>
<td>98</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Bachelor’s degree and higher</td>
<td>92</td>
<td>83</td>
<td>99</td>
<td>85</td>
<td>81</td>
<td>83</td>
</tr>
</tbody>
</table>

<sup>a</sup>X1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G1.

<sup>b</sup>X2: cognitive questionnaire correct rate of judgment and recognition part 1 for G1.

<sup>c</sup>X3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G1.

<sup>d</sup>Y1: cognitive questionnaire overall correct response rate of judgment and recognition of the COVID-19 health rumor for G2.

<sup>e</sup>Y2: cognitive questionnaire correct rate of judgment and recognition part 1 for G2.

<sup>f</sup>Y3: cognitive questionnaire correct rate of judgment and identification of the COVID-19 health rumor part 2 for G2.

**Figure 15.** Comparison between serious game learning mode and traditional learning mode.

**Discussion**

**Principal Findings**

Based on a self-made serious game, this paper investigated the health rumor phenomenon, and a study on the user behavior and willingness to disseminate health information among Chinese elderly WeChat users (SNS) was conducted during the early COVID-19 pandemic. After a survey, participants were chosen, and a COVID-19 health rumor was selected as the study content and the experimental platform with the self-made game was established. The UTAUT2 model was upgraded by adding parameters, several hypotheses were proposed, and a control experiment was designed. The experiment results show that the serious game is useful for health rumor prevention.

After collecting game data and the correct response rates of G1 and G2 in the cognitive questionnaire for the judgment and recognition of the COVID-19 health rumor, the game data and the cognitive questionnaire data were combined to determine the relationship between specific variables. Finally, the
experimental hypotheses were tested and evaluated, proving that H1-H6, H8, and H10 hold, while H7 and H9 do not hold. The findings affirm that serious games are a powerful tool to enhance learning and commonsense against health rumors in the context of elderly users of SNS. As Wu [43] argued, perceptions of rumor credibility affect the users’ desire to find accurate information (cognitive gratification) because they use SNS for verifying the contents of rumors and for acquiring more knowledge and information. Equipping oneself with better knowledge and commonsense against health rumors could have a profound effect on the stability and harmony of society [58], minimize the chance of being misinformed [35], and help create effective control strategies against rumor spreading [45]. Furthermore, as serious games provide the means for people to receive direct feedback relative to their judgment of health rumors, using these games is considered a more humane and emotional approach [47]. In addition, it also provides a suitable channel for health care providers to increase awareness [49] since tackling COVID-19 requires everyone to follow medical advice. Based on the verification of our hypotheses, we found that the effect of serious games correlates with parameters such as education, which suggests that the future rumor management for the youth is perfectly suited to the use of serious games, especially in China, where the education level of the youth is much higher than that of the middle-aged and older populations.

Limitations
Given the seminal findings of this study, it has some limitations. First, the cognitive questionnaire was administered offline, and the midelder/elder participants were reluctant in terms of their willingness to cooperate with the research. As such, there was a risk of the sample distribution being uneven or biased. Second, strict epidemic prevention and control have geographically limited experimental samples. Third, the serious game design was restricted to the COVID-19 health information and had limited interactivity.

Conclusion
This experimental study on preventing new health rumors via serious games proves that the serious game learning mode can help research participants understand and learn about health and rumors. Furthermore, serious games make a more profound impression on people than traditional learning modes, while providing fertile ground for more comprehensive research in the future. In addition, serious games could provide suggestions and support in future research on rumor prevention and detection. In particular, the Chinese government ended the zero-COVID policy in December 2022, and many new health rumors related to the Omicron variant were found on the internet in China. This study could provide a method of challenging the new issue and the game could be updated for the current situation. More importantly, we discovered that serious games can act as an “informational vaccine” against rumors (if rumors are considered a kind of “informational virus or bacterium”), and in the future, we can conduct further research in this direction.

Acknowledgments
This study was supported by the Huazhong University of Science and Technology, Independent Innovation Fund (#450/5003450061).

Conflicts of Interest
None declared.

Multimedia Appendix 1
A priori questionnaire for health rumor recognition.
[DOCX File , 16 KB - games_v12i1e45546_app1.docx ]

References


20. Yi HL. Designing serious games to enhance political efficacy and critical thinking disposition for college students: the case of Taiwan. 2011 Presented at: 2011 Third International Conference on Games and Virtual Worlds for Serious Applications; May 4-6, 2011; Athens, Greece. [doi: 10.1109/vsgames.2011.29]


**Abbreviations**

GM: genetically modified  
HCW: health care worker  
IP: Internet Protocol  
SNS: social networking service  
TCP: Transmission Control Protocol  
UDP: User Datagram Protocol  
UTAUT: unified theory of acceptance and use of technology

©Shuo Xiong, Long Zuo, Qiwei Chen, Zhang Zeliang, Mohd Nor Akmal Khalid. Originally published in JMIR Serious Games (https://games.jmir.org), 26.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
Assessing the Importance of Content Versus Design for Successful Crowdfunding of Health Education Games: Online Survey Study

Hong Huang¹, PhD; Han Yu², PhD; Wanwan Li³, PhD

Corresponding Author:
Hong Huang, PhD

Abstract

Background: Health education games make health-related tasks enjoyable and interactive, thereby encouraging user participation. Entrepreneurs and health educators can leverage online crowdfunding platforms, such as Kickstarter, to transform their innovative ideas into funded projects.

Objective: This research focuses on health education game initiatives on Kickstarter. Through an online user survey, it aims to understand user perceptions and evaluate the significance of 8 distinct components that may influence the success of such crowdfunding initiatives.

Methods: A total of 75 participants evaluated games using 8 dimensions: game rules, learning objectives, narrative, content organization, motivation, interactivity, skill building, and assessment and feedback. The survey data were analyzed using descriptive statistical analysis, exploratory factor analysis, the Wilcoxon-Mann-Whitney test, and multivariate analysis.

Results: Exploratory data analysis showed that, among the 8 dimensions, skill building, content organization, and interactivity were the top-ranking dimensions most closely associated with crowdfunding health education game. The 8 dimensions can be grouped into 3 categories from the exploratory factor analysis: game content–, instruction–, and game design–related components. Further statistical analysis confirmed the correlation between these dimensions with the successful crowdfunding of health education games.

Conclusions: This empirical analysis identified critical factors for game proposal design that can increase the likelihood of securing crowdfunding support.

(JMIR Serious Games 2024;12:e39587) doi:10.2196/39587

KEYWORDS

game-based learning; rubrics; Kickstarter; learning game campaign; collaboration; user perception; design; health; learning; gaming; game; evaluation; organization; user; engagement; skill; feedback; assessment; analysis; correlation; crowdfunding; support

Introduction

Background

Digital strategies, particularly gamification, have introduced a refreshing dynamic to health education [1,2]. Platforms, such as Kickstarter [3], champion these tech-infused health games, providing a unique avenue for their development. By leveraging the power of crowdfunding, Kickstarter and similar platforms facilitate the evolution of health education games. This allows entrepreneurs, educators, game developers, and supporters to access essential resources and connect with audiences eager for meaningful health support and intervention.

Gamification in Health

Gamification in health integrates game-design elements into nongame health scenarios, aiming to boost user engagement and immersion in health solutions. This transforms routine health tasks into enjoyable, competitive activities. This approach leads to positive behavioral changes, improving overall health, fitness, and adherence to medical treatments and programs [1,2,4,5]. Gamification has been applied to a wide range of medical fields, including health education, medical therapy, obesity, and mental health [1,2,4,5].

Health education games are interactive digital tools specifically designed to impart knowledge or skills related to health and wellness [1,2]. These games transform traditional health-related lessons into enjoyable and engaging tasks, aiming to enhance retention and application of health information in daily life [1,2,4]. Serious health games, created primarily for specific health objectives rather than solely for entertainment, use gaming components to create an educational environment [1,2]. They use gaming components to facilitate a teaching
environment, enabling users to learn specific health skills or gain valuable health-related information [1,2]. Especially beneficial for long-term health and chronic-related applications, these games foster positive emotional or empathetic connections among users, leading to improved medical treatment plans and behavior changes [1,2,4].

**The Role of Crowdfunding in Promoting Health Education Games**

Given the modest initial investment required and the scale of crowdfunding, it is advocated that crowdfunding serves as a primary method to promote and support the development of health education games. With the recent success of platforms such as Kickstarter, researchers and health care advocates are turning to these tools to fund their projects [6,7]. Through crowdfunding, health educators, entrepreneurs, and other stakeholders can conduct their work to meet community needs while also achieving financial and community outreach goals. This method attracts a varied group of participants who contribute financially, participate in the development, and offer social support [8-12].

Health education games, similar to other game-based learning tools, motivate users by making health-related tasks more enjoyable [1,13]. Online crowdfunding can assist entrepreneurs and health educators with limited resources to translate their innovative ideas into solid and appealing content and formats [14,15]. Crowdfunding platforms help individuals transform ideas into fundable and actionable projects [16,17].

Crowdfunding for health education games benefits users’ self-efficacy, well-being, chronic disease management, and physical activity [9-11]. Rewards, feedback, and socialization elements are frequently used to gamify eHealth in crowdfunding-based health education games. Furthermore, health education games can positively change their health behaviors, benefiting their overall health and wellness [2]. Successful health education crowdfunding projects elicit both intrinsic (altruistic) and extrinsic (rewards and feedback) motivation in order to attract a diverse range of crowdfunding donors, and they work by effortlessly facilitating online digital health engagement [18]. This study aims to explore 8 critical evaluation dimensions from the user’s perspective that influence the success of crowdfunding campaigns for health education games. The findings will guide practitioners and entrepreneurs in strategizing and designing impactful crowdfunding campaigns for health education games.

**Related Works**

To understand the intricacies of successful crowdfunding for health education games, we performed a literature review to acquire insights on the various dimensions related to the subject. The literature review enabled us to systematically explore the dynamics of crowdfunding, the principles of game-based learning, and the factors that influence the success of health education games.

**Dynamics and Success Factors of Crowdfunding Initiatives**

To develop and promote content for successful crowdfunding campaigns, extensive planning, outreach, and marketing are required. Data suggest that the most popular crowdfunding projects are those that are creative, participatory, or consumable, such as games, technology, film and video, and art and design [19]. In general, crowdfunding projects have small funding sizes and offer various donor incentives, small gifts, or awards, which leads to a higher success rate for the projects [19]. Such success not only mirrors financial objectives but also nurtures the emergence of communities with shared interests [20]. Numerous game developers have used crowdfunding to fund the initial investment in educational applications [21]. This then encourages more entrepreneurs to participate in collaborative crowdfunding platforms and launch their projects.

Unlike a traditional purchase, crowdfunding involves a high level of social capital influence, particularly the status and reach on social network sites [17]. Social capital creates an online environment that combines collective knowledge, appeal, and emotional responses, enabling investors to make well-informed decisions [17]. This investment process shapes perception and investment behavior. The interaction mechanism has a broader and more pervasive contextual impact, and the crowdfunding campaign design and features also influence decisions [21].

Crowdfunding initiatives require both content richness and ownership diversity [22]. Several studies have explored strategies to optimize the success of such crowdfunding efforts [22,23]. Notably, during crowdfunding, potential investors often evaluate founders based on their personal communication skills and presentation, both of which influence investment decisions [24]. In addition, the use of specific language, the length of campaign text, the frequency of updates, and the inclusion of video in campaign texts have all been correlated with the success of crowdfunding campaigns [25,26]. Reducing the cognitive effort needed to understand campaign content has been shown to result in increased funding [15].

Researchers have also linked crowdfunding success to the trustworthiness and reputation of developers, as well as their experiences on social crowdfunding networks [27,28]. However, the quality of the presented information also plays an important role in determining crowdfunding success [9,29,30]. Factors that contribute to successful crowdfunding factors include the content of the campaign, audience participation, and the timing of fundraising development [31].

**Health Education Game Development and User Experience**

Gamification has been proven to enhance medication and treatment adherence among patients with chronic disease [4,32]. Health serious games, on the other hand, have been praised for their ability to help people with chronic illnesses improve their behavior [2,33]. These games mirror real-life challenges, allowing players to develop coping strategies [17]. They educate players about their condition and the necessary lifestyle alterations, with compelling storylines that ensure better engagement [15,17]. Game interactivity allows players to make
decisions, learn from outcomes, and receive feedback on health implications [2,17].

When evaluating the feasibility of a game proposal, it is important to consider both the organization and narrative of the content, as well as the effectiveness of interactive games as a learning tool [34]. A well-organized and clearly written proposal can help the investor understand the purpose, goals, and potential value of the project [31]. Interactive health games can educate users with content and skills [34]. Users can also actively engage with the material, explore and experiment with different concepts and strategies, and receive immediate feedback on their progress [35]. This can help them understand and retain the content and skills being taught.

Game rules and interactivity stand as important components in health game design. Game rules ensure alignment with educational objectives, and the inherent challenge-reward system in these games drives players to continue, thereby continuously learning and adopting healthier behaviors [5,13]. Defining game rules or challenges and delivering feedback can increase users’ self-concept, efficacy, knowledge skills, communication, and social support, resulting in better health behaviors for self-care and adherence, lowering health costs, and establishing a stronger health system [18].

Health education game users are drawn to characters that resemble them, experiencing validation when such characters are featured in media [36]. Young role models, especially those in media genres such as cartoons and video games, are particularly valued by these users [37]. For example, the motivation and design of the interactive health game series can focus on using positive role models to inspire and motivate players [37,38]. These role models are described as being successful in their adventures while also managing their health, which could help users, including children with chronic illnesses such as asthma or diabetes, feel more positive about their own abilities to manage their health and self-care [37,38].

Interestingly, health game players without specific medical conditions are often less certain about in-game decisions compared to their peers with those conditions [37]. Health education games allow players to try new things, fail, learn, and eventually win. Such games also motivate users to adopt a healthier lifestyle, adhere to medical advice when unwell, navigate life crises, and foster close social connections for support [39].

Regarding assessments and feedback mechanisms, health learners who receive personalized feedback and engage deeply with medical content tend to experience great benefit. This approach is especially effective in reaching younger individuals who might not typically consult other media or seek expert health advice [40]. Interactive health games not only foster communication and social support but also empower users to discuss their health with friends, family, and health care professionals. They also motivate users to actively seek out advice and support [37]. For instance, in a series of interactive health games, players accessed factual details about the causes, treatments, social contexts, and self-care options related to specific health topics [37].

Game-Based Learning Principles

One of the game-based learning principles that allows users to benefit from the game is the development of problem-solving skills [41], and educational games can assist users in developing these skills [41,42]. The modalties of game content representation should be adjusted to boost motivation and performance [43]. If learners cannot understand the app’s content, no matter how rich and useful it is or how beautiful the design is, the app’s entire instructional value is lost [44]. Learners can learn problem-solving, strategic and analytical thinking, decision-making, and other 21st century skills in narrative-centered learning environments [45].

Based on the constructivist learning theory, individuals gain deeper insights about the world through direct experiences and interactions [46,47]. Games offer a dynamic and interactive environment that aligns with this theory, enabling learners to actively explore, experiment, and tackle challenges [46,47]. The appeal of a game’s narrative indicates its potential to captivate users [48]. The game creators should focus more on the content, storyline, and interaction components of the game to attract individual users when determining whether it will be successful or not [48].

The quality of a learning game is significantly influenced by the effectiveness of user feedback [49,50]. Numerous studies have shown that feedback enhances learning outcomes [51]. It provides learners with clarity on their strengths and areas that need improvement; it also serves as a motivational tool, encouraging continuous learning even within the gaming context [51].

Educational games can customize learning experiences by gauging a student’s readiness, providing constructive feedback, and modifying the level of challenge [52]. It is essential for an educational game to have well-defined learning objectives that detail the desired skills and knowledge [53]. Game rules facilitate learning by allowing players to interact with their environment [54]. Achieving these objectives depends on adhering to specific rules, which may involve certain challenges or conditions that the learner must satisfy [43].

A learner’s level of motivation can greatly influence their enthusiasm or indifference toward a task [55,56]. Moreover, there is substantial evidence suggesting that motivation enhances cognitive functions, particularly influencing what learners focus on and how they assimilate information [57-59].

Literature suggests that multiple factors influence the success of crowdfunding campaigns, especially those related to health education games [51,54]. These range from the trustworthiness of the developers and the quality of information presented to the design and content of the game itself. Although previous studies have shed light on the general principles of game-based learning and the dynamics of crowdfunding, there remains a gap in understanding how these principles specifically apply to health education games on platforms such as Kickstarter. Moreover, the user’s perspective, which is crucial in determining the success of such campaigns, has not been thoroughly explored. We aim to bridge this knowledge gap by focusing on the user’s perception and evaluating the critical components.
that resonate most with potential users, thereby influencing the success of health education game initiatives on crowdfunding platforms.

**Objectives**

This study aims to provide a comprehensive overview of 17 health education game projects launched on the crowdfunding platform Kickstarter and to understand user perceptions concerning the important factors that determine the success of such health education game crowdfunding initiatives. To achieve this, we conducted a user survey using a health education assessment rubric specifically designed to evaluate the key components contributing to the success of these projects on Kickstarter.

### Methods

**Data Collection for Health Education Games**

A comprehensive keyword search using “Health, Education, Learning, Game” was conducted in August 2019 on Kickstarter, which identified 17 online health education game projects (Table 1). On the Kickstarter site, the system marked a project as “Successful” if it met or exceeded its financial goal within the time set by the creators. Conversely, projects that failed to meet their financial target within the designated period were labeled as “Unsuccessful” (Table 1).
Table. Descriptive data of health education game projects from the crowdfunding site Kickstarter. A project’s success on Kickstarter was determined by its ability to achieve its financial goal within the set time frame.

<table>
<thead>
<tr>
<th>Health education game</th>
<th>Pledge (US $)</th>
<th>Goal (US $)</th>
<th>Backer count, n</th>
<th>Country</th>
<th>Successful(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Playout: The Exercise Card Game [60]</td>
<td>11,011</td>
<td>10,000</td>
<td>224</td>
<td>United States</td>
<td>Yes</td>
</tr>
<tr>
<td>ACLS MegaCode Simulator for health care professionals [61]</td>
<td>328</td>
<td>1997</td>
<td>8</td>
<td>Canada</td>
<td>No</td>
</tr>
<tr>
<td>Blush by Renaissance [62]</td>
<td>5065</td>
<td>3500</td>
<td>80</td>
<td>Canada</td>
<td>Yes</td>
</tr>
<tr>
<td>Body Cycle Health Education App [63]</td>
<td>1778</td>
<td>20,000</td>
<td>41</td>
<td>United States</td>
<td>No</td>
</tr>
<tr>
<td>CHiLD - a psychological 2D RPG [64]</td>
<td>1199</td>
<td>554</td>
<td>92</td>
<td>Norway</td>
<td>Yes</td>
</tr>
<tr>
<td>Destiny’s Sword for mental health [65]</td>
<td>30,930</td>
<td>30,000</td>
<td>209</td>
<td>Canada</td>
<td>Yes</td>
</tr>
<tr>
<td>Facing Dragons: a mixed-reality game to unlock your purpose [66]</td>
<td>3361</td>
<td>7104</td>
<td>34</td>
<td>Canada</td>
<td>No</td>
</tr>
<tr>
<td>Freestyle Jam Camp [67]</td>
<td>1145</td>
<td>500</td>
<td>18</td>
<td>United States</td>
<td>Yes</td>
</tr>
<tr>
<td>Mobile games to quantify symptoms of mental health disorders [68]</td>
<td>127</td>
<td>450,000</td>
<td>5</td>
<td>United States</td>
<td>No</td>
</tr>
<tr>
<td>PRESCRIPTION Playing Cards [69]</td>
<td>30,420</td>
<td>7500</td>
<td>178</td>
<td>Canada</td>
<td>Yes</td>
</tr>
<tr>
<td>Talk to Me visual novel: mental health [70]</td>
<td>4977</td>
<td>4460</td>
<td>146</td>
<td>United States</td>
<td>Yes</td>
</tr>
<tr>
<td>TREN: a card game designed to promote brain health [71]</td>
<td>1445</td>
<td>14,000</td>
<td>39</td>
<td>United States</td>
<td>No</td>
</tr>
<tr>
<td>The Chakra Collectable Coin [72]</td>
<td>1682</td>
<td>1300</td>
<td>41</td>
<td>United States</td>
<td>Yes</td>
</tr>
<tr>
<td>The Woosah Kit: a mental health first aid [73]</td>
<td>41</td>
<td>6236</td>
<td>3</td>
<td>United Kingdom</td>
<td>No</td>
</tr>
<tr>
<td>Tournesol Kids Game: activity cards to build resilience [74]</td>
<td>10,435</td>
<td>5000</td>
<td>140</td>
<td>United States</td>
<td>Yes</td>
</tr>
<tr>
<td>Youth Run The World 5K [75]</td>
<td>7370</td>
<td>7000</td>
<td>74</td>
<td>United States</td>
<td>Yes</td>
</tr>
<tr>
<td>Zombied: gamify health and fitness activities [76]</td>
<td>12</td>
<td>37,217</td>
<td>2</td>
<td>United Kingdom</td>
<td>No</td>
</tr>
</tbody>
</table>

\(^a\)“Yes” refers to “Successful” projects that met or exceeded their financial goal, whereas “No” refers to “Unsuccessful” projects that did not.

Ethical Considerations

Before commencing this study, the researchers obtained approval from the Institutional Review Board of the University of South Florida (001588). The participants provided informed consent, with the option to withdraw at any time without penalty. The Institutional Review Board approval sufficiently covered the secondary use of data. The study guaranteed that all collected data were either anonymized or deidentified to protect personal information, with stringent protective measures in place for any data that could not be fully anonymized. The study was voluntary, without any compensation for participation.

Online Survey Design

We use the Qualtrics online survey platform (Qualtrics) to create an online survey based on health education game assessment rubrics derived from the literature. This survey allowed participants to evaluate and rank crowdfunding health education projects.
games on the Kickstarter website. The survey incorporated 8 dimensions—each essential for the evaluation of health education games. These dimensions, along with their definitions and cited literature, are presented in Table 2.

Table. Crowdfunding health education game evaluation dimensions and definitions.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Definition</th>
<th>Related literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skill building</td>
<td>The game’s ability to progressively impart and reinforce health-related skills to players, ensuring that learning is continuous and effective throughout the game’s duration.</td>
<td>[42,77]</td>
</tr>
<tr>
<td>Content organization</td>
<td>The clarity, structure, and logical flow of the game’s health education content, ensuring that it is presented in a manner that is both comprehensible and engaging for players.</td>
<td>[35,53,78]</td>
</tr>
<tr>
<td>Narrative</td>
<td>The clarity and continuity of the game’s storyline in relation to health education, ensuring that players experience a coherent sense of progression and purpose as they navigate through the game’s content.</td>
<td>[25,48]</td>
</tr>
<tr>
<td>Interactivity</td>
<td>The game’s ability to facilitate effective interactions, the completion of health-related tasks, and active participation through user-driven inputs and actions.</td>
<td>[35,77,79]</td>
</tr>
<tr>
<td>Assessment and feedback</td>
<td>The game’s capability to immediately evaluate and communicate a player’s progression and provide timely and relevant feedback.</td>
<td>[35,80-82]</td>
</tr>
<tr>
<td>Game rules</td>
<td>The game provides clear, concise, and easily comprehensible rules to the players.</td>
<td>[35,52,83,84]</td>
</tr>
<tr>
<td>Learning objectives</td>
<td>The game delineates specific, measurable outcomes that players are anticipated to achieve upon its completion.</td>
<td>[35,85-87]</td>
</tr>
<tr>
<td>Motivation</td>
<td>The game’s elements are intriguing and appealing enough to prompt user participation and action.</td>
<td>[88-90]</td>
</tr>
</tbody>
</table>

Before the main survey was launched, a pilot test of the survey instrument was conducted with 7 undergraduate students majoring in health science. This pilot test aimed to assess the validity and understandability of the survey questions. The participants were asked to read through the survey and provide feedback on its clarity and relevance. Based on their comments, necessary revisions were made to the questions to enhance the overall quality of the survey.

In the final version of the survey, participants rated the dimensions on a 3-point Likert scale. The scoring system for these dimensions ranged from 0 to 2, with the following interpretations: 0=“Does not meet expectations” or “Poor,” 1=“Meets expectations” or “Fair,” and 2=“Exceeds expectations” or “Good.” Participants could also select “Unable to decide” or “Not applicable” if they felt unable to make a judgment on a particular dimension. Additionally, an open-ended question was incorporated: “Do you have any comments or concerns (accuracy of terms, comprehensiveness, clarity of questions, etc) for this question sets?” This allowed participants to provide further feedback on the survey questions.

In November 2019, undergraduate students majoring in health science were invited to participate in the online survey. Those who agreed to participate were provided with a standardized set of questions, accompanied by comprehensive instructions and definitions for the 8 evaluation dimensions, as detailed in Table 2. Each student was then randomly assigned 1 specific crowdfunding health education game from a pool of 17 games, referenced in Table 1. Their task was to evaluate their assigned game based on these 8 dimensions. Ultimately, 75 undergraduate students were recruited as participants.

Data Analysis

We used STATA 15 software (StataCorp) for statistical analyses. We used several data analysis approaches to understand the results.

Descriptive Statistical Analysis

This method provides a summary of the main aspects of the data, offering a simple overview of the data. By calculating the percentage of ranking types and the mean scores of the dimensions, we can gain insights into the general behavior and preferences of the survey participants.

Exploratory Factor Analysis

Exploratory factor analysis is used to reduce the data’s dimensionality and identify the underlying relationships between the measured variables [91]. It was used to group the 8 dimensions into meaningful categories, helping to decipher any latent structures within the data set. This ensured that we could identify which sets of dimensions tended to co-occur or were rated similarly by participants.
Wilcoxon-Mann-Whitney Test

The Wilcoxon-Mann-Whitney test [92] is a nonparametric statistical hypothesis test used to compare 2 unrelated samples. This test was used to determine if there were any significant differences in the rankings given by participants to different game dimensions.

Multivariate Analysis

The aim of this study extends beyond merely understanding the dimensions. It also seeks to predict the success of crowdfunding health education games based on these dimensions. We used logistic regression with a binary variable—success of the crowdfunding project—for prediction [91]. This model can determine the odds of a game being successful based on the rankings of its dimensions, offering insights into which dimensions are the most influential predictors of success.

By using these methods, the study ensured a comprehensive analysis of the data—from understanding the basic patterns and deciphering underlying component structures to finally being able to predict the success of crowdfunding health education games based on their dimensions.

Results

A list of health education games launched on Kickstarter is presented in Table 1. This table enumerates 17 distinct health education games originating from various countries, namely the United States, Canada, Norway, and the United Kingdom. Some projects have exceeded their goals by a large margin, whereas others have fallen substantially short. The diversity of the sample provides a comprehensive foundation for our study. This diversity enabled an exploration into users’ perceptions regarding educational game assessment rubrics. Such an investigation can discern potential factors that could influence the success trajectory of health education games on crowdfunding platforms such as Kickstarter.

Table 2 focuses on the various dimensions relevant to the design and evaluation of games. These dimensions were based on established literature, highlighting their credibility and validity. When assessing potential predictors of crowdfunding success based on feedback from 75 survey participants, certain dimensions stood out as being more important (Table 3).

Table 1. Ranking of the 8 assessed dimensions for crowdfunding health education games (n=75).

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skill building</td>
<td>1.77 (0.54)</td>
</tr>
<tr>
<td>Content organization</td>
<td>1.7 (0.52)</td>
</tr>
<tr>
<td>Narrative</td>
<td>1.51 (0.69)</td>
</tr>
<tr>
<td>Interactivity</td>
<td>1.51 (0.75)</td>
</tr>
<tr>
<td>Assessment and feedback</td>
<td>1.49 (0.69)</td>
</tr>
<tr>
<td>Game rules</td>
<td>1.47 (0.71)</td>
</tr>
<tr>
<td>Learning objectives</td>
<td>1.39 (0.64)</td>
</tr>
<tr>
<td>Motivation</td>
<td>1.29 (0.59)</td>
</tr>
</tbody>
</table>

aScoring system: 0="Poor," 1="Fair," and 2="Good."

Skill building was ranked first, followed by content organization and then narrative. Skill building holds the top rank due to its emphasis on continuous learning and engagement, ensuring that players progressively acquire and refine their skills throughout the game (Table 3). The importance of content organization is highlighted by its role in enhancing user experience; a well-organized game offers clear navigation, allowing players to immerse themselves fully (Table 3). Narrative further enhances the gaming experience by introducing an engaging storyline that lends context and purpose, enriching the gameplay. Interactivity is important for keeping players engaged. It gives them a sense of belonging and influence within the game world. Yet, intriguingly, motivation ranks the lowest among these dimensions, even though its presence ensures that games are compelling enough to retain players’ interest and drive continuous participation (Table 3). Although skill building and content organization seem to be the areas where these games excel, motivation appears to be a challenging area for many developers.

To identify the assessment structure for campaign initiatives’ quality reflected by 75 survey respondents’ rankings, the study conducted an exploratory factor analysis using principal-components analysis as the extraction method and varimax with Kaiser normalization as the rotation method (Table 4). The cutoff size for criterion loadings was set to 0.45 [59]. Both the Bartlett ($\chi^2=68.26, P<.001$) and measure of sampling adequacy (0.57) tests for the sample pointed to a significant level of correlation among the dimensions.
Table 1. Factor components for the 8 dimensions in crowdfunding health education games. Principal-components analysis served as the extraction method, and varimax with Kaiser normalization served as the rotation method.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game rules</td>
<td>−0.050</td>
<td>−0.057</td>
<td>0.843&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Learning objectives</td>
<td>0.253</td>
<td>0.727&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−0.025</td>
</tr>
<tr>
<td>Narrative</td>
<td>0.181</td>
<td>0.629&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−0.335</td>
</tr>
<tr>
<td>Motivation</td>
<td>0.665&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.102</td>
<td>−0.021</td>
</tr>
<tr>
<td>Interactivity</td>
<td>0.489</td>
<td>0.185</td>
<td>0.506&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Skill building</td>
<td>−0.136</td>
<td>0.727&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.266</td>
</tr>
<tr>
<td>Assessment and feedback</td>
<td>0.883&lt;sup&gt;a&lt;/sup&gt;</td>
<td>−0.036</td>
<td>−0.019</td>
</tr>
<tr>
<td>Content organization</td>
<td>0.490&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.370</td>
<td>0.033</td>
</tr>
</tbody>
</table>

<sup>a</sup>Values above the cutoff size for criterion loadings (0.45).

The exploratory factor analysis indicated that these 8 dimensions can be grouped into 3 components: game content (content organization, motivation, and assessment and feedback), instruction (learning objectives, narrative, and skill building), and game design (game rules and interactivity; Table 4). The game content–related components suggests that a well-organized game with clear feedback mechanisms can effectively motivate players. The instruction-related components reflect the instructional journey of the player, from understanding the objectives and engaging with narrative to building skills. The game design–related components are fundamental to the gameplay experience, ensuring that players are not just passive observers but active participants.

To review the perception gaps among these dimensions for successful or unsuccessful crowdfunding campaigns, group-based comparison was conducted between these dimensional means. Table 5 showed the gaps between successful and unsuccessful games in dimension ratings. Among them, motivation, interactivity, game rules, and learning objectives demonstrated larger difference gaps in decreasing order, and these were followed by assessment and feedback, skill building, narrative, and content organization.
Table . Wilcoxon-Mann-Whitney test of the 8 assessments based on successful or unsuccessful crowdfunding of health education games.

<table>
<thead>
<tr>
<th>Dimensions and categories</th>
<th>Answers, n</th>
<th>Score, mean (SD)</th>
<th>U statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content organization</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.68 (0.55)</td>
<td>0.28</td>
<td>.78</td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>15</td>
<td>1.67 (0.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interactivity</strong></td>
<td></td>
<td></td>
<td>2.05</td>
<td>.04</td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.57 (0.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>15</td>
<td>1.13 (0.83)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Skill building</strong></td>
<td></td>
<td></td>
<td>0.94</td>
<td>.35</td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.79 (0.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>15</td>
<td>1.60 (0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Learning objectives</strong></td>
<td></td>
<td></td>
<td>2.03</td>
<td>.04</td>
</tr>
<tr>
<td>Success</td>
<td>51</td>
<td>1.43 (0.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>15</td>
<td>1.07 (0.59)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Narrative</strong></td>
<td></td>
<td></td>
<td>.09</td>
<td>.37</td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.53 (0.70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>14</td>
<td>1.36 (0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Motivation</strong></td>
<td></td>
<td></td>
<td>2.91</td>
<td>.004</td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.38 (0.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>15</td>
<td>0.87 (0.52)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Game rules</strong></td>
<td></td>
<td></td>
<td>2.14</td>
<td>.03</td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.55 (0.70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>15</td>
<td>1.13 (0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assessment and feedback</strong></td>
<td></td>
<td></td>
<td>1</td>
<td>.32</td>
</tr>
<tr>
<td>Success</td>
<td>53</td>
<td>1.49 (0.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsuccessful</td>
<td>14</td>
<td>1.29 (0.73)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant level P<.05.

The Wilcoxon-Mann-Whitney test comparing distributions of successful and unsuccessful games showed that motivation (P=.004), game rules (P=.03), learning objectives (P=.04), and interactivity (P=.04) showed statistically significant difference among these 2 groups (Table 5). These dimensions showed clear distinctions between successful and unsuccessful games, suggesting that these dimensions might be crucial for the success of such games. On the other hand, dimensions such as content organization and skill building, while important, did not show a significant difference between the 2 categories of games. This could mean that both successful and unsuccessful games have well implemented these dimensions, but they might not be the distinguishing factors for success. The multivariate analysis showed that learning objectives and motivation were 2 significant dimensions associated with successful health education game crowdfunding campaigns (Table 6). This suggests that these 2 dimensions might be especially important for the success of health-related games.
Table. Multivariate logistic regression predicting the success of the health educational games.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Odds ratio (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game rules</td>
<td>3.24 (0.68-18.40)</td>
<td>.13</td>
</tr>
<tr>
<td>Learning objectives</td>
<td>3.55 (1.42-14.38)</td>
<td>.02*</td>
</tr>
<tr>
<td>Narrative</td>
<td>1.57 (0.37-6.61)</td>
<td>.54</td>
</tr>
<tr>
<td>Content organization</td>
<td>0.07 (0.01-1.55)</td>
<td>.09</td>
</tr>
<tr>
<td>Motivation</td>
<td>3.05 (1.46-9.36)</td>
<td>.03*</td>
</tr>
<tr>
<td>Interactivity</td>
<td>1.70 (0.46-6.22)</td>
<td>.42</td>
</tr>
<tr>
<td>Skill building</td>
<td>1.31 (0.21-8.18)</td>
<td>.77</td>
</tr>
<tr>
<td>Assessment and feedback</td>
<td>1.38 (0.21-8.88)</td>
<td>.73</td>
</tr>
</tbody>
</table>

*Significant level P<.05.

Figure 1 presents an empirical framework that outlines the key components underpinning the success of health education game crowdfunding. The model highlights the balance between foundational structural components, such as game rules and content organization, and experiential elements that enhance the player’s immersion and engagement, such as motivation and narrative. A successful educational game should seamlessly integrate all these facets. This not only ensures the delivery of educational content but also fosters an environment where players are intrinsically driven to remain engaged and continue their learning journey within the game.
Discussion

Principal Findings

The crowdfunding landscape for health education games is diverse, with success determined by a myriad of factors beyond just a funding goal. Factors such as the clarity of the project’s purpose, its presentation, and its marketing likely play a substantial role in attracting users [35,41]. It is also important to have a reasonable and attainable goal, as this might increase the likelihood of a project’s success.

Crowdfunding backers, especially on platforms such as Kickstarter, often support projects that offer value beyond just entertainment. Skill building in games implies that players will acquire new abilities or knowledge, making them both fun and beneficial. This dual-purpose might appeal to game players who see an opportunity for a return on investment, not just in potential product rewards but also in personal or societal skill development.

The ranking of these dimensions sheds light on the preferences and priorities of both backers and players. It is possible that backers perceive tangible attributes such as skill building and content organization as immediate indicators of game quality and potential success. These elements can be readily demonstrated in promotional materials, making them more attractive to potential backers. On the other hand, motivation, being more abstract and subjective, might be harder to convey and measure, leading to its lower ranking. It is essential for
game developers to recognize these perceptions and strike a balance in their design, ensuring a comprehensive and engaging game experience that appeals to a broad audience.

For skill building, it is essential for players to acquire and build skills as they progress in the game. This ensures continuous learning and engagement. Well-structured game content helps players navigate and understand the game better, thus enhancing their experience. An engaging storyline provides context and purpose, making gameplay more meaningful. Player interactivity is vital for player engagement. Players should feel that they are part of the game world and can influence it. Immediate feedback helps players understand their progression and areas of improvement. Clear rules ensure that players can easily understand how to play the games, leading to smoother game experiences. For health education games, it is important to have clear learning outcomes that guide the game design. The game must be engaging enough to keep players interested and motivated to continue.

The multivariate analysis identified learning objectives and motivation as the 2 significant predictors of a health education game’s crowdfunding success, as detailed in Table 6. This indicates the emphasis users place on clear educational outcomes and the motivation to engage with the game. Users prioritize games that offer clear educational outcomes and that effectively motivate players to engage. The significance of learning objectives suggests that backers might prioritize games that have a clear educational goal, ensuring that players gain tangible knowledge or skills. Motivation, on the other hand, ensures that players remain engaged and committed to the game’s objectives. When combined, these dimensions can lead to a game that not only educates but does so in a compelling manner, maximizing player retention and learning outcomes.

Limitations and Future Work
The study has some limitations due to the examination of user perception, which is based on a small number of user responses in a small number of crowdfunding campaigns. The study examined subjective opinions across 8 evaluation dimensions, but the reasons for crowdfunding’s effectiveness in health education games require further investigation. In addition, we surveyed participants as potential backers. A more comprehensive approach would involve surveying actual backers, those who make real investments, to discern any differences in perceptions. This could provide a richer understanding of the dynamics at play. The impact of quality on the campaign content and media aspects, as well as user indicators of motivation and interactivity, was investigated in this study. Through crowdfunding, health education games improve engagements, learning components, and cultural adaptability for user engagement [8-10].

Conclusion
Crowdfunding for health education games presents a unique opportunity to bridge the gap between game developers and potential users. There has been little research that has provided empirical evidence for evaluating user perspectives on crowdfunding health education games. Further empirical evaluations are clearly beneficial to providing a rigorous validation of gamification’s effectiveness in eHealth. This research conducted an exploratory study and identified 3 major components that matter for health game crowdfunding success. These components are related to game design, instruction, and game content. Interestingly, motivation and assessment and feedback were grouped into game content categories, not into game design categories. This indicates that the proposals for health-related crowdfunding education games are comprehensive, encompassing content that is engaging, interesting, and attractive, with solid assessment and feedback components. Among them, given the nature of health subjects, entrepreneurs and educators should pay more attention to game development factors such as motivation, interactivity, and game rules, so that the health or scientific subjects can be easily infused in the gaming process. Making health games look playful and attractive enables users to easily grasp basic health knowledge during the gaming process [93]. Interestingly, there is little difference in content organization between successful and unsuccessful games, which indicates that even if the game content is easy to follow, it is still not enough. Backers and potential funders or users mostly agree with the health content itself, but they care more about the game development components, using these dimensions to assess the crowdfunding game proposal and determine if these game designs are acceptable and make logical sense.

Our findings recognize the importance of aligning game design with user preferences. The success of health education games on crowdfunding platforms relies on a combination of clear educational objectives, effective player engagement mechanism, and well-structured game content. The study highlights the significance of learning objectives and motivation as key determinants of crowdfunding success for health education games. Game developers aiming for success in this domain should prioritize these dimensions, thus ensuring that their games offer a clear educational outcome.

Acknowledgments
We would like to extend our sincere gratitude to the 3 anonymous reviewers for their invaluable and insightful feedback on our manuscript. Their consistent comments significantly improved our manuscript.

Conflicts of Interest
None declared.

References

https://games.jmir.org/2024/1/e39587


43. Reeves TC, Harmon SW. Systematic evaluation procedures for instructional hypermedia/multimedia. Presented at: Annual Meeting of the American Educational Research Association; Apr 14, 1993; Atlanta, GA.


Abstract

Background: Artificial intelligence (AI) and game-based methods such as serious games or gamification are both emerging technologies and methodologies in health care. The merging of the two could provide greater advantages, particularly in the field of therapeutic interventions in medicine.

Objective: This scoping review sought to generate an overview of the currently existing literature on the connection of AI and game-based approaches in health care. The primary objectives were to cluster studies by disease and health topic addressed, level of care, and AI or games technology.

Methods: For this scoping review, the databases PubMed, Scopus, IEEE Xplore, Cochrane Library, and PubPsych were comprehensively searched on February 2, 2022. Two independent authors conducted the screening process using Rayyan software (Rayyan Systems Inc). Only original studies published in English since 1992 were eligible for inclusion. The studies had to involve aspects of therapy or education in medicine and the use of AI in combination with game-based approaches. Each publication was coded for basic characteristics, including the population, intervention, comparison, and outcomes (PICO) criteria; the level of evidence; the disease and health issue; the level of care; the game variant; the AI technology; and the function type. Inductive coding was used to identify the patterns, themes, and categories in the data. Individual codings were analyzed and summarized narratively.

Results: A total of 16 papers met all inclusion criteria. Most of the studies (10/16, 63%) were conducted in disease rehabilitation, tackling motion impairment (e.g., after stroke or trauma). Another cluster of studies (3/16, 19%) was found in the detection and rehabilitation of cognitive impairment. Machine learning was the main AI technology applied and serious games the main game-based approach used. However, direct interaction between the technologies occurred only in 3 (19%) of the 16 studies. The included studies all show very limited quality evidence. From the patients' and healthy individuals' perspective, generally high usability, motivation, and satisfaction were found.

Conclusions: The review shows limited quality of evidence for the combination of AI and games in health care. Most of the included studies were nonrandomized pilot studies with few participants (14/16, 88%). This leads to a high risk for a range of biases and limits overall conclusions. However, the first results present a broad scope of possible applications, especially in motion and cognitive impairment, as well as positive perceptions by patients. In future, the development of adaptive game designs with direct interaction between AI and games seems promising and should be a topic for future reviews.

(JMIR Serious Games 2024;12:e48258) doi:10.2196/48258
KEYWORDS
artificial intelligence; AI; games; serious games; gamification; health care; review

Introduction

Background

Artificial intelligence (AI) and serious games are both relevant topics in the health sector, and the body of studies and literature is continuously growing. Interestingly, in terms of the research landscape, the 2 topics are not connected; rather, existing research views them independently.

The use of games for educational and serious purposes is nearly as old as the history of humankind and is an integral part of our culture [1]. In 1970, Abt [2] used the term “serious games” for the first time in his book with the same name. Sawyer and Smith [3] take a broad definition and consider serious games as “any computerized game whose chief mission is not entertainment and all entertainment games which can be reapplied to a different mission other than entertainment.” What serious games have in common is that they pursue a concrete (pedagogical) intention and aim to provide information on a specific topic (eg, health) that is accessible in an entertaining and interactive way to deepen competencies or to achieve a change in behavior [4].

Serious games for health can be used in the fields of medical diagnostics, therapy, and prevention, as well as health promotion and medical or patient education [5]. From a didactic and learning psychology perspective, the effect of serious games is based on the integration of the motivating and multimedia aspects of computer and video games. Serious games can increase engagement, motivation, enthusiasm, and interest [6,7]. There are several existing use cases in health contexts [8-11]. One example is the game EndeavorRx. In 2020, the US Food and Drug Administration permitted its marketing as the first game-based digital therapeutic device to improve attention function in children with attention-deficit/hyperactivity disorder (ADHD) [12]. The game Re-Mission was developed for children with cancer and showed good results regarding compliance and the understanding of disease-related issues in the target group [13]. EMERGE is a simulation game that recreates an emergency department in real time to improve the clinical reasoning skills of physicians [14].

Next to serious games, gamification has emerged as a major trend in the health sector, which is reflected in a growing number of publications, including several meta-analyses [15-17]. The most used definition of this concept is “the use of game design elements in non-game contexts” [18]. The motivational effect of the game elements can be explained in different ways. Sailer et al [19] established the link between various gamification elements (eg, points, leaderboards, and badges) and the self-determination theory proposed by Ryan and Deci [20]. As a theory of motivation, this defines three universal psychological basic needs that determine human action: (1) competence, (2) autonomy, and (3) social inclusion. If ≥1 of these needs are addressed (eg, through gamification elements), this has positive effects on behavior and its determinants [19]. In the health sector, there are numerous studies that have demonstrated the effects of using gamification on motivation, performance, engagement, health, and well-being status [5,21,22].

According to Westera et al [23], computer games have been linked with AI since the first computer was programmed to play chess [24]. New AI methods have been used in computer games, for instance, to generate levels, scenarios, and storylines; to balance complexity; or to add intelligent behaviors to nonplayer characters (NPCs) [25]. However, over the years, various authors have pointed at the marginal penetration of academic game AI methods in industrial game production [26]. AI techniques will become indispensable to coordinate the ever-growing complexity and dynamics of games [23]. AI-driven adaptation and assessment systems are used to offer learner-centered environments [27]. As an example, NPCs controlled by AI can adapt to the behavior of the gamer and can enrich immersive and challenging experiences within the game play.

When transferring these principles to health care, the interaction between AI and games could provide a benefit, especially in the management of chronic diseases, which most game designs already target. The possibility to quickly adapt to new game-generated data or performance and provide live feedback could lead to more individual and thus more patient-centered game design in both illness detection and treatment. This could increase motivation and engagement for patients, leading to higher therapy adherence through more personal involvement. The vast body of evidence in the field of serious games and gamification, along with the growing body of evidence in the use of AI, may thus form a new field of research.

Scope

Therefore, this scoping review sought to generate an overview of the currently existing literature on the interaction of AI and game-based approaches in health care. At this point, to the best of our knowledge, this is the only review that targets this interaction.

The primary objective was to analyze the current body of evidence based on (1) the disease or health issue being evaluated, (2) the process of care in which these projects are located, and (3) the kind of AI and type of game-based approach used and the interaction of both techniques.

A secondary objective was to obtain an overview of publications on the interaction of AI and game-based approaches such as serious games, gamification, commercial games, and game periphery. Another secondary objective was to analyze the quality of the existing studies in this field regarding their grade of evidence and the conducted study types.

Methods

Overview

For this scoping review, we applied the PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols; Figure 1) guidelines [28]. Furthermore, we used the
recommendations of the Cochrane Consortium for conducting systematic reviews and the RefHunter website as guidance [29].

Before starting the review process, we defined the inclusion and exclusion criteria (Textbox 1).

**Figure 1.** PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) flow diagram.

**Textbox 1.** Inclusion and exclusion criteria.

**Inclusion criteria**
- Article type: original study, journal article, or conference paper
- Article scope: articles report the use of artificial intelligence (AI), machine learning, and deep learning in combination with game-based approaches (serious games, gamification, and game-based-learning)
- Health profession: medicine
- Area of application: articles that conducted research in the field of education, therapy, and health
- Language: English
- Publication period: last 20 years

**Exclusion criteria**
- Article type: opinion, commentary, or letter to the editor
- Article scope: not related to AI and game-based approaches
- Health profession: other than medicine
- Area of application: not related to health and medicine
- Language: not in English
- Publication period: published >20 years ago

For this review, we conducted the following steps:

1. Literature search
2. Title and abstract screening
3. Content screening
4. Further in-depth screening (snowballing method and asking colleagues)

**Step 1: Literature Search**
We applied the search terms primarily in the database PubMed on February 2, 2022. We tested and honed different search terms and Boolean operators (Multimedia Appendix 1) until sufficiently fitting results seemed to have been obtained (n=305). The final search term was defined as follows:

[“game” OR “gamification”] AND “artificial intelligence”
The search was extended to more open databases to assess studies that target AI and serious games in medicine-related research areas or interprofessional approaches that may include medical professions and in more technically oriented databases such as IEEE Xplore [30] to include papers from informatics and engineering with a focus on technical issues.

The same search term ("game” OR “gamification”) AND “artificial intelligence”) AND “artificial intelligence”) was used for IEEE Xplore (n=98), Cochrane Library [31] (n=25), and PubMed Psych [32] (n=89). In Scopus, the search term used in the other databases showed fewer results and were modified to extend the range of hits ("serious” AND “game” AND “artificial intelligence”; n=41).

In addition, we conducted a search in Google Scholar [33]. However, the results from Google Scholar were not precise enough for inclusion in a review, which is consistent with the results of several studies [34-36].

After deduplication using Rayyan software (Rayyan Systems Inc), the combined search in these databases identified 545 (97.7%) publications out of the initial 558 identified. We then performed a manual deduplication, which resulted in 60 (11%) of the 545 publications being excluded; thus 485 (89%) publications remained (Multimedia Appendices 1 and 2; Multimedia Appendix 3 [37-51]).

**Step 2: Title and Abstract Screening**

For the second step of the scoping review (title and abstract screening), we used Rayyan software. The results from the literature search were transferred to the citation software Zotero (version 5.0.85; Rayyan) and to Rayyan software [52]. This software automatically identified duplicates. After iterative deduplication, the publications were subjected to manual screening. The first screening step was conducted using Rayyan and permitted publication inclusion based on their titles and abstracts. Given the volume of the publications to be screened, the title and abstract screening was distributed among 2 authors of this paper (JS and DT). To ensure the uniformity of the screening, the authors conducted several training sessions in Rayyan with the coreviewers. In addition, the authors randomly double-checked some of the excluded publications (25/485, 5.2%) to warrant the consistency of the screening by the other reviewer. This step was conducted independently by both researchers and was followed by a discussion of the results between the 2. As all analysis steps were conducted independently by the 2 researchers, a discussion of differently categorized literature (marked as “Conflict” in Rayyan) and subsequent adaptation were necessary in this step. Overall, only a few adjustments were necessary, and a good agreement between the 2 researchers could be reached. Possible conflicts and all included articles were discussed with a third team member (SK). The data generated by Rayyan can be found in Multimedia Appendix 2.

**Step 3: Content Screening**

After the primary screening, full-text publications were screened by the 2 lead authors. Toward this end, a table was prepared to compile relevant information (Textbox 2).

### Textbox 2. Information compiled for full-text screening.

<table>
<thead>
<tr>
<th>Relevant details obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors, year, title, journal, and digital object identifier (DOI)</td>
</tr>
<tr>
<td>Study type (according to Röhrig et al [53])</td>
</tr>
<tr>
<td>Population, intervention, comparison, and outcomes (PICO) criteria</td>
</tr>
<tr>
<td>Subject (topic of the study)</td>
</tr>
<tr>
<td>Level of evidence (according to the Oxford Centre for Evidence-Based Medicine: Levels of Evidence [54])</td>
</tr>
<tr>
<td>Disease or health issue</td>
</tr>
<tr>
<td>Level of care (prevention, diagnostics, therapy, rehabilitation, nursing, organization or monitoring, and other [55])</td>
</tr>
<tr>
<td>Game variants (serious games, gamification, games, and game controller or periphery)</td>
</tr>
<tr>
<td>Artificial intelligence (AI) technology (machine learning, deep learning, and AI [not further specified])</td>
</tr>
<tr>
<td>Function type (promoting health literacy, analysis and cognition, indirect intervention, direct intervention, documentation of health and medical history, organization and administration, and purchasing and supply) [56]</td>
</tr>
</tbody>
</table>

Some of the studies (4/16, 25%) showed an overlap among different categories (eg, in level of care). In these cases, double classifications were performed. All eligible studies were categorized and coded in detail (Multimedia Appendix 3 [37-51]).

**Step 4: Further In-Depth Screening (Snowballing Method and Asking Colleagues)**

After conducting the scoping review, we additionally used the “snowballing” approach described by Greenhalgh and Peacock [57], who have stated that in reviews of complex and heterogeneous evidence, formal protocol-driven search strategies may fail to identify important evidence. Informal approaches such as browsing and asking colleagues can substantially increase the efficiency of search efforts. Snowballing methods such as pursuing references of references and electronic citation tracking are very useful for identifying high-quality sources in obscure locations. Therefore, to validate the results of the review, the 2 reviewers searched the literature references used in meta-analyses, reviews, and papers that were closely related to the topic of the search. In addition to using the snowballing...
method, the method of asking colleagues, as recommended by Greenhalgh and Peacock [57], was applied as a last step.

Results

Overview
When applying the aforementioned search terms, we initially identified 335 studies on the topic of games and the use of AI in health care in the last 20 years. The subsequently performed step of title and abstract screening reduced the number of the initially identified studies from 335 to 47 (14%). In the next step, assessing the actual full-text literature, of the 47 papers, 3 (6%) were excluded because their full text was not in English, and the aforementioned inclusion and exclusion criteria were applied to the remaining 44 (94%). After the full-text screening, 10 (23%) of the 44 papers met all inclusion criteria. Using the snowballing method, 1 additional paper could be identified. Asking colleagues revealed 5 additional papers, which led to an overall total of 16 eligible papers (Figure 1). Not all criteria showed hits (eg, function type showed hits only in 2 categories, whereas level of care showed no hits in nursing).

Categories

Overview
The eligible papers showed a clear emphasis on certain categories (Table 1). Regarding the targeted diseases, the field of motion impairment was investigated the most (5/16, 31%). Cognitive impairment was targeted in 19% (3/16) of the studies, phantom limb pain or limb absence in 19% (3/16), rheumatoid arthritis in 13% (2/16), cancer in 6% (1/16), and ADHD in 6% (1/16). The primary focus on rehabilitation (10/16, 63%) was the most compelling. Of the 16 studies, 5 (31%) took place in the field of prevention, 4 (25%) in the field of diagnostics, and 1 (6%) in a nonrehabilitation therapeutic context (4 double assignments).

Most of the studies (12/16, 75%) applied machine learning as the AI technology, and 13% (2/16) used deep learning, whereas the remaining studies (2/16, 13%) did not specify the AI technology. Most of the studies (11/16, 69%) used a serious game, whereas 19% (3/16) used a commercial games approach. Despite the highly increased use of gamification in health and education, of the 16 studies, only 1 (6%) specifically used gamification to improve the motivation of patients, and 1 (6%) used game design–like interactions.

We further clustered and outlined the eligible papers according to the targeted disease (Table 2). A more detailed description of every included publication with a more specific outline of the use of AI and game variant can be found in Multimedia Appendix 4 [37-51].
<table>
<thead>
<tr>
<th>Category</th>
<th>Studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease or health topic (n=16)</strong></td>
<td></td>
</tr>
<tr>
<td>Motion impairment</td>
<td>5 (31)</td>
</tr>
<tr>
<td>Phantom limb pain or limb absence</td>
<td>3 (19)</td>
</tr>
<tr>
<td>Cognitive impairment</td>
<td>3 (19)</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>2 (13)</td>
</tr>
<tr>
<td>Cancer</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Attention-deficit/hyperactivity disorder</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Function type (n=16)</strong></td>
<td></td>
</tr>
<tr>
<td>Direct intervention</td>
<td>9 (56)</td>
</tr>
<tr>
<td>Analysis and cognition</td>
<td>7 (44)</td>
</tr>
<tr>
<td><strong>Level of care (n=20(^a))</strong></td>
<td></td>
</tr>
<tr>
<td>Prevention</td>
<td>5 (25)</td>
</tr>
<tr>
<td>Diagnostics</td>
<td>4 (20)</td>
</tr>
<tr>
<td>Therapy</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>10 (50)</td>
</tr>
<tr>
<td><strong>AI(^b) technology (n=16)</strong></td>
<td></td>
</tr>
<tr>
<td>Machine learning</td>
<td>12 (75)</td>
</tr>
<tr>
<td>Deep learning</td>
<td>2 (13)</td>
</tr>
<tr>
<td>AI (not further specified)</td>
<td>2 (13)</td>
</tr>
<tr>
<td><strong>Game variant (n=16)</strong></td>
<td></td>
</tr>
<tr>
<td>Serious games</td>
<td>11 (69)</td>
</tr>
<tr>
<td>Gamification</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Games</td>
<td>3 (19)</td>
</tr>
<tr>
<td>Game periphery</td>
<td>1 (6)</td>
</tr>
</tbody>
</table>

\(^a\) A total of 4 studies showed an overlap between prevention and diagnostics and were double classified, resulting in an overall total of 20 studies.

\(^b\) AI: artificial intelligence.
Table 2. Overview of included papers, structured by disease or health topic.

<table>
<thead>
<tr>
<th>Disease or health topic</th>
<th>Game variant</th>
<th>Authors; year</th>
<th>Target group (participants, n)</th>
<th>Subject</th>
<th>Study design</th>
<th>Level of evidence</th>
<th>Level of care</th>
<th>Function type</th>
<th>AI technology</th>
<th>Game variant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease or health topic: motion impairment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yeh et al [43]; 2014</td>
<td>Patients (48)</td>
<td>Noninvasive balance training</td>
<td>Case control study</td>
<td>3b</td>
<td>Therapy</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lyu et al [44]; 2019</td>
<td>Healthy individuals (8)</td>
<td>Electromyography-controlled knee exoskeleton</td>
<td>Quantitative, proof of concept</td>
<td>5</td>
<td>Rehabilitation</td>
<td>Analysis and cognition</td>
<td>Deep learning</td>
<td>Games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nasri et al [45]; 2020</td>
<td>Patients (15)</td>
<td>Real-time hand gesture recognition</td>
<td>Case series</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Deep learning</td>
<td>Serious games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Burdea et al [42]; 2021</td>
<td>Healthy individuals (2)</td>
<td>Game controller–based telerehabilitation</td>
<td>Proof of concept</td>
<td>5</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>AI</td>
<td>Game controller or periphery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Zhang et al [46]; 2021</td>
<td>Patients (5)</td>
<td>Gait analysis and waist motion capture</td>
<td>Case series</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td><strong>Disease or health topic: phantom limb pain or limb absence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ortiz-Catalan et al [47]; 2016</td>
<td>Patients (14)</td>
<td>Phantom motor execution</td>
<td>Quantitative clinical trial</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lendaro et al [48]; 2019</td>
<td>Patients (4)</td>
<td>Phantom motor execution</td>
<td>Quantitative clinical trial</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Analysis and cognition</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kristofferson et al [49]; 2021</td>
<td>Patients (4)</td>
<td>Prosthesis system</td>
<td>Explorative study</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td><strong>Disease or health topic: cognitive impairment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Valladares-Rodriguez et al [50]; 2018</td>
<td>Healthy individuals or patients (16)</td>
<td>Early detection of mild cognitive impairment</td>
<td>Proof of concept</td>
<td>5</td>
<td>Prevention or diagnostics</td>
<td>Analysis and cognition</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Valladares-Rodriguez et al [51]; 2019</td>
<td>Patients (74)</td>
<td>Early detection of mild cognitive impairment</td>
<td>Case series</td>
<td>4</td>
<td>Prevention or diagnostics</td>
<td>Analysis and cognition</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jung et al [37]; 2019</td>
<td>Patients (12)</td>
<td>Mini-Mental State Examination</td>
<td>Case series</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td><strong>Disease or health topic: cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Good et al [39]; 2014</td>
<td>Registered players (1077)</td>
<td>Gene Selection for breast cancer survival prediction</td>
<td>Quantitative study</td>
<td>5</td>
<td>Prevention</td>
<td>Analysis and cognition</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td><strong>Disease or health topic: attention-deficit/hyperactivity disorder</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keshav et al [40]; 2019</td>
<td>Patients (7)</td>
<td>Digital attention-related augmented reality game</td>
<td>Case series</td>
<td>4</td>
<td>Prevention or diagnostics</td>
<td>Analysis and cognition</td>
<td>AI</td>
<td>Serious games</td>
</tr>
<tr>
<td><strong>Disease or health topic: rheumatoid arthritis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Varga et al [38]; 2021</td>
<td>Healthy individuals (7)</td>
<td>Virtual arthritis rehabilitation app</td>
<td>Proof of concept</td>
<td>5</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Varga et al [58]; 2022</td>
<td>Patients (10)</td>
<td>Virtual arthritis rehabilitation app</td>
<td>Case series</td>
<td>4</td>
<td>Rehabilitation</td>
<td>Direct intervention</td>
<td>Machine learning</td>
<td>Serious games</td>
</tr>
</tbody>
</table>
Almost one-third of the studies (5/16, 31%) targeted the objective of motion impairment. Studies included upper- and lower-limb rehabilitation with a broad range of possible medical indications, ranging from poststroke to vestibular dysfunction. Games were used to enhance motivation and provide a user-friendly at-home training experience. Some of the studies (5/16, 31%) achieved this through an integration of virtual reality and artificial reality. AI was integrated in different ways. Some of the studies (4/16, 25%) used games as a training tool and then analyzed and classified the collected data with AI. Other studies (3/16, 19%) first processed sensor data via AI to improve the quality of an associated game. Direct interaction between the AI and the games component was shown in 2 (40%) of the 5 studies, in which AI adapted the game design and difficulty to the ability level of the patient.

Only 1 (20%) of the 5 studies tested the design using a control group analyzing patient improvements in clinical parameters. All other studies demonstrated the functionality and usability of their technical approach in pilot studies.

Phantom Limb Pain or Limb Absence

Of the 16 studies, 3 (19%) targeted the topic of phantom limb pain or limb absence, where a game environment can support at-home therapy and provide enhanced visual feedback. Of the 3 studies, 2 (67%) by the same research group targeted phantom motor execution with similar approaches. Machine learning was used to improve the quality of electromyography sensor data and thus provide better data input for training. Different training methods in the spectrum of virtual reality and augmented reality and serious games were tested. Of the 3 studies, 1 (33%) focused on ethnographic user–type analysis, and 1 (33%) effected a decrease in phantom pain. The third study tested a machine learning–aided prosthesis, comparing 2 different training approaches—one conventional and 1 via a serious game—to collect electromyography data. Testing was only conducted on 4 patients; however, the results were insignificant.

Cognitive Impairment

In cognitive impairment, the included studies used a set of games covering different cognitive functions as diagnostic instruments. Data were then processed by machine learning techniques to further improve outcome quality. Of the 3 studies, 1 (33%) focused on evaluating patients with cognitive impairment after stroke. Scores acquired from a game set were analyzed by AI and compared with the clinically widely used Mini-Mental State Examination (MMSE) [37]. Of the 3 studies, 2 (67%) used a game set for predicting the future development of mild cognitive impairment, using AI to automatically distinguish between healthy individuals and individuals who were possibly affected. In both fields, pilot studies were conducted with patients, showing high motivation to participate and good usability of the game sets.

Rheumatoid Arthritis

In rheumatoid arthritis, a serious game for hand rehabilitation was developed. Neural networks for processing data and machine learning for testing the accuracy of hand movements for individually adapting difficulty were integrated. Two small pilot studies, 1 with healthy individuals and 1 with patients, showed high accuracy of the machine learning algorithm and good usability, whereas clinical benefits have not yet been measured [38].

Cancer

In cancer, a crowdsourcing campaign was set up via an open web-based game that captured inputs from players regarding their estimation of 5 specific genes, which can be used as predictors of breast cancer survival. Gene selections were processed by machine learning to identify the best prediction models. When only including inputs from people with a self-proclaimed Doctor of Medicine degree, a Doctor of Philosophy degree, or expertise in cancer, the resulting models performed similarly to clinically established gene sets [39].

ADHD Severity

A set of smartglasses was developed to assess ADHD severity through playing an attention-related augmented reality game designed as a social-emotional communication aid. AI was used to analyze video and audio as well as affective and behavioral data and provided users with in-game rewards based on their performance. The study showed significant correlation of the game score to validated clinical gold standard assessments for ADHD [40].

Other

To improve the prevention of cognitive and physical decline, an at-home innovative system consisting of remote monitoring and neurocognitive games was developed. Feedback to the user, including badges or benefits for real-life events, is provided via machine learning analysis. Older adult users indicated “great acceptability” of the system [41].

---

**Table: Disease or health topic: other**

<table>
<thead>
<tr>
<th>Authors; year</th>
<th>Target group (participants, n)</th>
<th>Subject</th>
<th>Study design</th>
<th>Level of evidence&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Level of care</th>
<th>Function type</th>
<th>AI&lt;sup&gt;b&lt;/sup&gt; technology</th>
<th>Game variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinto et al [41]: 2019</td>
<td>Older adults (11)</td>
<td>Active and assisted living for monitoring daily life activities</td>
<td>Case series</td>
<td>4</td>
<td>Prevention or diagnostics</td>
<td>Analysis and cognition</td>
<td>Machine learning</td>
<td>Gamification</td>
</tr>
</tbody>
</table>

<sup>a</sup>According to the Oxford Centre for Evidence-Based Medicine: Levels of Evidence [54].

<sup>b</sup>AI: artificial intelligence.
Discussion

Principal Findings
Currently, there are only a limited number of studies involving a combination of game-based methods and AI in health. Almost one-third of the included studies (5/16, 31%) were centered on addressing motion impairment. The primary emphasis of the research was on rehabilitation. In addition, most of the studies (9/16, 56%) focused on prevention and diagnostics. In terms of AI technology, machine learning was the most commonly used approach (12/16, 75%). Furthermore, serious games were used in most of the studies (11/16, 69%).

When analyzing the studies by disease category, most of the studies (5/16, 31%) used a rehabilitation approach for different aspects of motion impairment (eg, in poststroke conditions, phantom limb pain or limb absence, and rheumatoid arthritis). In this field, studies have a focus on providing individual, at-home, and complex training opportunities for improving motoric limb function, in which therapeutic concepts rely on long-term and self-guided exercising. Games take the role of a training tool, enhancing at-home training motivation and providing multidimensional and exercises compared with the current standard of care. In addition, in some of the studies (5/16, 31%), the integration of virtual reality and augmented reality provided an immersive experience. The role of AI in this context is diverse, sometimes to analyze and classify collected data to improve game setup and level, sometimes to analyze data resulting from game play itself.

In a second cluster, studies for neurological diseases, including those handling cognitive impairment in older adults as well as 1 study for ADHD in younger patients, there was a clear focus on diagnostic evaluations. Here, different sets of games were used to assess various cognitive subdomains, with AI processing these different data inputs and calculating scores and predictions. The advantages in this field are the wide range of possible game designs and the feasibility to play these games individually at home. This could reduce health professionals’ time in assessing cognitive function during face-to-face visits or supplement them by enabling longitudinally acquired data sets and trajectories. The first results show promising results in comparison with standard clinical scores obtained using, for instance, the MMSE.

The direct interaction between the games approach and AI technology was only described in 3 (19%) of the 16 studies. Most of the time, the 2 entities follow each other, with the AI technology not analyzing live in-game playing data. However, direct interaction holds a promise of benefit through an AI-enabled assessment of the patient’s ability during game play and individualized live adjustments of game design and difficulty. Examples using this approach showed good technical functioning and positive user feedback [42]. Even so, the limited number of published studies suggest that the potential of this integrated approach has so far not been fully used yet. This is rather surprising, given the fact that the direct link between AI and games is widely prevalent in the commercial games sector.

The reasons for this are purely speculative. The transition of findings from 1 field to another is still pending, perhaps because studies in the commercial games field have a different scope than those in medicine and health. Another reason could be the resource-intensive nature of research. However, the future potential of this interaction seems promising, with the stimulation of user motivation by game design and gamification elements and with AI being used to process large and multimodal data sources and to perform individualized adaptations.

When analyzing further categories, our review shows that the studies so far have produced very limited quality evidence (all studies have an evidence level of 4 or 5, except for 1 study that has an evidence level of 3b), with most of the studies presenting either a rather technical proof of concept (15/16, 94%) or performing usability testing with a small sample size of healthy individuals and patients (14/16, 88%). Higher-quality studies with control groups and end points focusing on specific clinical outcomes are missing.

Of note, the research field is still young. All studies were conducted in the last 8 years, with 13 (81%) of the 16 studies being published in the last 3 years. All research settings however bear the potential of conducting higher-quality studies with bigger sample sizes and specific medical outcomes in the near future.

However, the studies in this review already show promising results, with overall well-functioning technical implementation of the game elements and high accuracy and usefulness of the AI integration. From the patients’ and healthy individuals’ perspective, generally high usability, motivation, and satisfaction were found, mostly assessed by established usability questionnaires and qualitative interviews. This is an encouraging perspective for the future because individualized patient-driven at-home diagnostic and therapeutic approaches are increasingly relevant in all fields of medicine.

All 16 studies identified in this review have a relatively low level of evidence (3b: n=1, 6%; 4: n=10, 63%; and 5: n=5, 31%). The risks of bias in these studies are multifaceted. Pilot studies, often conducted to assess the feasibility of a full-scale study, typically featured small sample sizes and often lacked rigorous methodology, randomization, and blinding procedures. As a result, they are susceptible to a range of biases, including selection bias, performance bias, and detection bias. Studies were characterized by weaker methodologies, which can lead to biases in data collection, analysis, and reporting. Nonrandomized studies were prone to selection bias, confounding, and other methodological flaws. The high heterogeneity of the identified studies encompassed a wide range of disease or health issues, populations, and interventions. This heterogeneity makes it challenging or impossible to integrate data and limits overall conclusions.

Limitations
First, as described earlier, the field of research is still very interdisciplinary, and the studies carried out are very diverse based on the vast variety of game-based approaches and therapeutic interventions.

This review only covered original studies in English, which were found in the PubMed, Scopus, IEEE Xplore, Cochrane Library, and PubMed databases and published in the last 20
years. Although these are widely recognized and commonly used databases in the field of health care research, restricting the review to these 5 databases may have resulted in the exclusion of relevant studies published in other databases owing to this high interdisciplinarity. However, efforts have been made to minimize this limitation using comprehensive search strings, snowballing, and asking colleagues to identify additional relevant literature. In addition, this review also includes interdisciplinary databases such as the more technical-oriented IEEE Xplore and the more pedagogical-oriented PubPsych.

It especially remains unclear whether all projects conducted especially with a more technical focus have been published in scientific journals at all. For future reviews, a more holistic approach should be taken to assess more results from projects that may not have been included in a publication.

In addition, there might be a lack of awareness that research in the engineering, gaming, and fitness spectrum has a direct connection with health-related issues. Thus, it seems possible that certain publications were not fully covered by our already broad search strategy or that promising interventions have not been related to health care yet. This should be mitigated in future studies, considering the growing attention to this young research field.

Another limitation is that this review focused on therapeutic medical interventions rather than on health interventions. AI and game-based approaches in the field of prevention and health promotion have not been included, although this is an important aspect of population health. Game-based approaches especially are used a lot in this field to reach the target groups [8,9,21,22,59-61].

Future Directions

In the near future, the potential of games, which is already established in the commercial games sector, should be applied to the field of serious games and AI. Adaptive game design can be suitable in health care to improve the intervention outcome via AI-driven health care games that assess the skills level of the patient and adapt the difficulty in feedback loops, which could lead to a better harmonization with traditional therapy sessions. NPCs could be used as virtual patients or other health care personnel or relatives to simulate the interprofessional working environment and to improve the interaction and communication with virtual patients [26,27,34].

Finally, the integration of AI and games should carefully consider the ongoing discussions regarding ethical, moral, and data protection issues. In particular, studies describing ethical issues using game-based approaches are scarce [62,63].

Analyzing the currently limited evidence with promising future possibilities in study design and quality, as well as a dynamic research field, it seems, at this stage, that another review should be conducted in the next few years to assess this rapidly growing research field.

Acknowledgments

Open access funding was provided by the open access publishing funds of Philipps-University Marburg and Bielefeld University.

Authors' Contributions

All authors were responsible for conceptualization, methodology, validation, and visualization. DT and JJS were responsible for data curation and formal analysis. DT and SK were responsible for investigation. SK was responsible for funding acquisition, project administration, resources, software, and supervision. DT and JJS wrote the original draft; all authors reviewed and edited the manuscript (contributions of the authors are based on CRediT [contributor roles taxonomy] [64]).

Conflicts of Interest

SK is the founder and a shareholder of MED.digital. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Search term protocol.
[DOCX File , 31 KB - games_v12i1e48258_app1.docx ]

Multimedia Appendix 2
Rayyan data of the review process of all studies.
[XLSX File (Microsoft Excel File), 320 KB - games_v12i1e48258_app2.xlsx ]

Multimedia Appendix 3
Categorization of all eligible studies.
[XLSX File (Microsoft Excel File), 32 KB - games_v12i1e48258_app3.xlsx ]

Multimedia Appendix 4
Detailed descriptions of all included studies.
[DOCX File , 25 KB - games_v12i1e48258_app4.docx ]

https://games.jmir.org/2024/1/e48258
References


30. IEEE Xplore. URL: https://ieeexplore.ieee.org/Xplore/home.jsp [accessed 2023-10-10]


32. PubPsych. URL: https://www.pubpsych.eu/ [accessed 2023-10-10]

33. Google Scholar. URL: https://scholar.google.com/ [accessed 2023-12-20]


35. Giustini D, Boulus MN. Google Scholar is not enough to be used alone for systematic reviews. Online J Public Health Inform 2013 Jul 01;5(2):214 [FREE Full text] [doi: 10.5210/ophi.v5i2.4623] [Medline: 23923099]


54. OCEBM levels of evidence. Centre for Evidence-Based Medicine (CEBM), University of Oxford. URL: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/cebm-levels-of-evidence/ [accessed 2023-12-20]


64. CRediT - contributor roles taxonomy. CRediT. URL: https://credit.niso.org/ [accessed 2023-10-11]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
AI: artificial intelligence
MMSE: Mini-Mental State Examination
NPC: nonplayer character
Tolks D, Schmidt JJ, Kuhn S
The Role of AI in Serious Games and Gamification for Health: Scoping Review
JMIR Serious Games 2024;12:e48258
URL: https://games.jmir.org/2024/1/e48258
doi: 10.2196/48258
PMID: 38224472

©Daniel Tolks, Johannes Jeremy Schmidt, Sebastian Kuhn. Originally published in JMIR Serious Games (https://games.jmir.org), 15.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
The Effects of Serious Games on Cardiopulmonary Resuscitation Training and Education: Systematic Review With Meta-Analysis of Randomized Controlled Trials

Pengfei Cheng, MSc; Yangxi Huang, MSc; Pengyu Yang, MSc; Haizhen Wang, MSc; Baichao Xu, PhD; Chaoran Qu, MSc; Hua Zhang, PhD

1Department of Nursing, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China
2School of Nursing, The University of Hong Kong, Hong Kong, China
3Department of Nursing, West China Hospital, Sichuan University, Chengdu, China
4Department of Physical Education, Hainan Medical University, Haikou, China
5Department of the Operating Room, Shenzhen People’s Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, China
6International Nursing School, Hainan Medical University, Haikou, China
7Key Laboratory of Emergency and Trauma, Ministry of Education, Haikou, China

*these authors contributed equally

Corresponding Author:
Hua Zhang, PhD
International Nursing School
Hainan Medical University
Number 3, Xueyuan Road.
Haikou, 571199
China
Phone: 86 13698954876
Email: zhanghuashelle@hotmail.com

Abstract

**Background:** Serious games have emerged as an innovative educational strategy with the potential to significantly enhance the quality and effectiveness of cardiopulmonary resuscitation (CPR) training. Despite their promise, there remains a degree of controversy when comparing the advantages of serious games with traditional CPR training methods. This study seeks to provide a comprehensive assessment of the impact of serious games on CPR training and education by systematically analyzing the results of previous research.

**Objective:** This study aimed to assess the effect of serious games on CPR training and education by summarizing and pooling the results of previous studies.

**Methods:** We conducted a thorough and systematic search across 9 prominent web-based databases, encompassing the period from the inception of these databases until April 1, 2023. The databases included in our search were PubMed, Cochrane Library, Wiley Online Library, EBSCO (PsycInfo), SpringerLink, Chinese Biology Medicine Disc, Vip Journal Integration Platform, Wanfang Database, and Chinese National Knowledge Infrastructure. The studies selected adhered to the following criteria: (1) being a randomized controlled trial comparing serious games and traditional methods for CPR training; (2) having participants aged 12 years or older in CPR; (3) having an experimental group using serious games and a control group using nongame methods for CPR instruction; and (4) having outcomes including theoretical and skill assessments, compression depth, and rate. The Cochrane risk of bias assessment tool was used to evaluate the risk of bias. Data analysis was performed using RevMan (version 5.3; Cochrane Training), and mean differences (MDs) and standardized mean differences (SMDs) with 95% CIs were used to calculate continuous variables.

**Results:** A total of 9 articles were included, involving 791 study participants, of whom 395 in the experimental group taught CPR training using serious games and 396 in the control group taught CPR training using traditional methods. The results of our meta-analysis indicate that the use of serious games in CPR training yields outcomes that are comparable in effectiveness to traditional training methods across several key areas. Specifically, serious games demonstrated equivalence to traditional formats
in theory assessment (SMD = −0.22, 95% CI = 0.96 to 0.51; P = .55), skill assessment (SMD = −0.49, 95% CI = −1.52 to 0.55; P = .36), compression depth (MD = −3.17, 95% CI = −1.80 to 6.53; P = .64), and compression rate (MD = −0.20, 95% CI = −7.29 to 6.89; P = .96).

Conclusions: In summary, serious games offer a viable and effective CPR education approach, yielding results comparable to traditional formats. This modality is a valuable addition to CPR training methodologies. However, caution is warranted in interpreting these findings due to limited controlled trials, small sample sizes, and low-quality meta-analyzed evidence.

(JMIR Serious Games 2024;12:e52990) doi:10.2196/52990

KEYWORDS
CPR; education; meta-analysis; serious game; training

Introduction

Background

Out-of-hospital cardiac arrest (OHCA) is a critical medical emergency characterized by the sudden cessation of heart function, resulting in an abrupt loss of blood flow. OHCA incidents frequently occur in community settings, schools, homes, and public places [1]. Despite sustained efforts, OHCA survival rates remain disheartening, largely due to modifiable factors such as bystander cardiopulmonary resuscitation (CPR), automated external defibrillator (AED) use, and the timing of emergency medical services (EMS) intervention [2,3]. In the United States, OHCA affects over 88.8 adults per 100,000 adults annually, with a mere 9.0% discharge survival rate, as reported by the American Heart Association [4]. Similarly, in Europe, the annual incidence of OHCA among adults ranges from 67 to 170 per 100,000, with discharge survival rates varying from 0% to 18% [5]. In China, more than 540,000 individuals experience OHCA each year, but the survival rate remains at approximately 2% [6]. These statistics underscore that OHCA, despite regional disparities, has emerged as a substantial public health challenge, imperiling the well-being of citizens [7]. OHCA is typified by sudden respiratory distress, pulse cessation, and loss of consciousness, necessitating immediate and effective first-aid measures within the critical 4-minute window [8]. However, current prehospital EMS services often struggle to reach the scene promptly to address emergencies in public spaces [9]. Consequently, first responders (FRs), nonmedical professionals in public areas, shoulder the responsibility of on-site rescue efforts [10]. Swift and efficient basic life support interventions administered by FRs not only create a vital time buffer for EMS teams to arrive but also substantially elevate the chances of patients with OHCA surviving [11].

CPR, encompassing artificial respiration and chest compressions, stands as one of the simplest and most universally applicable techniques for basic life support during OHCA emergencies [12]. The quality of chest compressions holds immense significance in preserving organ perfusion. Consequently, the timely and effective administration of CPR plays a pivotal role in determining both the survival rate and neurological outcomes for patients with OHCA [3]. To enhance the widespread adoption of CPR and ensure that more individuals are proficient in this vital first-aid technique, the World Health Organization and the International Liaison Committee on Resuscitation endorsed the “Kids Save Lives” statement, which calls for CPR training for students, adolescents, and adults aged 12 years or older who already have the physical fitness and learning ability to understand and remember CPR skills to empower young people, including children aged 12 years, with CPR skills. Develop a generation of proactive and empowered community members who are expected to make a difference in emergency situations, especially in the context of OHCA, with the goal of increasing survival and improving long-term outcomes for patients with OHCA [13,14].

Serious games are increasingly used in medical education, encompassing medical theory instruction, clinical skills training, cognitive rehabilitation exercises, and patient health education. The integration of serious games into medical simulation programs is seen as a means to enhance the efficiency and effectiveness of training programs [15,16]. Otero-Agra et al [17] used serious games to instruct middle school students in CPR, revealing that 61.7% of participants acquired correct CPR techniques, with 93.4% achieving an average chest compression depth exceeding 50 mm. These results endorse serious games as effective tools for knowledge acquisition and the mastery of high-quality CPR skills. To optimize their use as an educational strategy, serious games must possess robust content and cater to the target audience. Integrating learning theory with game requirements enhances student engagement and ensures the efficacy of learning [18]. High fidelity is crucial, especially for medical students, as the knowledge and skills acquired in serious games will be applied in future clinical practice involving real patients. High-fidelity serious games bridge the gap between virtual gaming scenarios and clinical reality, boosting rescue confidence and self-efficacy [19]. Creutzfeldt et al [20] used serious games based on massively multiplayer virtual worlds technology to train 36 high school students in CPR. After 90-120 minutes of game-based sessions, participants reported a significant increase in self-efficacy, endorsing the effectiveness of serious games for CPR instruction. Moreover, serious games can incorporate adaptive learning features, adjusting difficulty and content based on the learner’s proficiency, ensuring tailored learning for individuals with varying CPR skill levels [21].

The incorporation of serious games into CPR training aims to enhance the learning process by rendering it more engaging, interactive, and effective. Compared to conventional methods relying on lectures, videos, and hands-on practice, serious games make the learning experience more enjoyable, interactive, and motivation-driven, integrating features such as scores, levels, and rewards [21,22]. Notably, serious games for CPR training are user-facing, offering immediate training opportunities, flexible learning schedules, and detailed real-time feedback on CPR performance [23]. In contrast, traditional teaching models often limit training opportunities, providing delayed feedback,
particular in large-scale group activities where individual feedback is frequently overlooked [24]. A systematic review by Lim et al [25] underlines that the absence of regular retraining and effective feedback in traditional CPR education can impact skill retention. Serious games address these shortcomings by providing continuous opportunities for practice and feedback. Moreover, serious games support collaborative learning, enabling learners to respond jointly to virtual CPR scenarios and develop teamwork and communication skills. They also offer diverse immersive first-aid scenarios with varying causes of cardiac arrest, an aspect unattainable in traditional teaching formats [16,26]. This multifaceted approach not only compensates for the deficiencies in traditional methods but also promotes a dynamic and engaging learning environment in CPR training. Considering the advantages mentioned above, the 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care recommended the incorporation of serious games into CPR training and education to enhance teaching methods and improve instructional quality, taking into account advancements in training equipment and teaching formats [27]. However, Dankbaar [28] concluded that serious games have limitations in terms of time and their ability to provide learners with sufficient knowledge acquisition and complex skill improvement. In summary, there exists a degree of controversy regarding the impact of serious games on CPR training and education. Therefore, we aimed to conduct a meta-analysis to determine the effectiveness of serious games in CPR training and education.

Research Gap and Aim

While numerous researchers have explored and experimented with serious games for CPR training, published randomized controlled trial (RCT) studies have explored and experimented with the effect of serious games applied to CPR training, and their effectiveness has been proven and supported [16,29]. However, due to the limitations of research, the generalization of research conclusions is affected. Specifically, (1) these RCTs were single-center studies with small sample sizes; (2) specific serious games limit the reliability of the findings in different settings of serious games or target populations; (3) outcomes were mostly assessed by questionnaires, and there were a lack of reliable, automated, and repeatable methods to measure their efficacy; and (4) there is a lack of methodological specifications and standard protocols for the use of serious games. Furthermore, there is a lack of systematic evaluation or meta-analysis of the effectiveness of serious games-based CPR training. Consequently, it is necessary to quantitatively analyze the objective effect of serious games–based training through meta-analysis. In view of this, we conducted a meta-analysis to comprehensively evaluate the effect of serious games on CPR training and teaching.

Methods

Overview and Registration

This systematic review adheres to the guidelines set forth by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [30] and was registered in advance in the PROSPERO (International Prospective Register of Systematic Reviews) database (registration number CRD42023423089).

Search Strategy

Our search was conducted in several databases, including PubMed, Cochrane Library, Wiley Online Library, EBSCO (PsycInfo), and SpringerLink. Besides, Chinese databases, including the China National Knowledge Infrastructure (CNKI), China Biomedical Literature Database, VIP Journal Integration Platform, and Wanfang Database, were searched. The search was conducted from the inception of the databases until April 1, 2023. We limited the publication language to English and Chinese. English search terms included “serious game,” “gam*,” “cardiopulmonary resuscitation,” “CPR,” “basic life support,” “BLS,” “first aid training,” “resuscitation education,” “emergency skill,” etc. The search involved a combination of subject terms and free words, with a manual retrospective search of references and associated literature to ensure a comprehensive search of relevant studies. Multimedia Appendix 1 provides detailed information on the search strategies, including search terms, and the process used.

Eligibility Criteria for This Review

The eligibility of studies was assessed based on the following criteria: (1) the study type should be an RCT comparing the effectiveness of serious games with other traditional training methods for teaching CPR; (2) the study population should include participants aged 12 years or older who participated in CPR training or first aid training that covered basic life support for CPR; (3) interventions in the experimental group should involve the use of serious games for CPR training instruction, while the control group should receive other methods of CPR theory and skills training instruction excluding serious games, with no limitations on the types of games or software used; and (4) outcome measures should include one or more of the indicators of theoretical assessment, CPR skill assessment, compression depth, and compression rate. Additionally, duplicate or multiple manuscripts, literature in languages other than Chinese or English, literature with inaccessible full text, incomplete or missing data, improper data collection, or errors in statistical methods were excluded.

Screening Process

Two authors (PC and PY), who were trained in evidence-based methods, independently conducted the screening of literature and extraction of data. All references were managed using EndNote X9 (Clarivate), a reference management software. After removing duplicates, the remaining references were first screened based on titles and abstracts. Subsequently, full-text screening was performed independently by the authors in duplicate to determine the inclusion of literature. Disagreements were resolved through discussion or adjudication by a third author (HZ).

Quality Assessment

The Cochrane handbook’s criteria for assessing the risk of bias in RCTs were used to evaluate the methodological quality of the trials [31]. The assessment covered various aspects, including selection bias, concealment of the allocation scheme, implementation bias, measurement bias, missed visit bias,
reporting bias, and other biases. Each item was categorized as “low risk of bias,” “unclear,” or “high risk of bias.” In cases where differing opinions arose, a third author (HZ) was involved to reach a consensus.

**Data Extraction**

For data extraction, we used Excel (2010; Microsoft Corporation) to create a standardized form. The form included the following information: (1) basic details such as the first author, publication year, and country of the study; (2) population characteristics, sample size, and information about the serious games used in training and teaching; (3) specific interventions for the test and control groups; and (4) outcome measures and the tools used for measurement.

**Statistical Analysis**

Data analysis was carried out using RevMan (version 5.3; Cochrane Training). To assess heterogeneity, the Q test and the $I^2$ test were used. If the resulting $P$ value was greater than or equal to .1 and $I^2$ was less than or equal to 50%, it indicated low heterogeneity among the findings, leading to the selection of the fixed-effects model for meta-analysis. Otherwise, the random-effects model was used. When comparing groups, continuous variables were analyzed using mean difference (MD) if the same measurement instrument was used or standardized mean difference (SMD) if different instruments were used. Both effect measures were reported with 95% CIs. For continuous data that did not follow a normal distribution in the included studies and were expressed as medians, extreme values, or quartiles, a specific web-based formula calculator developed by Professor Luo et al. [32] from Hong Kong Baptist University was used. This calculator, designed for meta-analysis data conversion, enabled statistical estimation of the data. Leave-one-out analysis was used to conduct sensitivity analysis, that is, omitting one study at a time from the meta-analysis and examining the impact on the overall effect size, then judging the robustness and reliability of the results and exploring the sources of heterogeneity [33]. Statistical significance was determined at a $P<.05$. The level of evidence was evaluated using the GRADEpro GDT web-based tool.

**Results**

**Study Selection**

After conducting a comprehensive search across various databases, a total of 843 RCTs were found. Additionally, 7 more studies were obtained by snowballing. Following the removal of duplicates, 415 articles were screened based on their titles and abstracts. Out of these, 382 articles were excluded, and the remaining 33 articles were examined in their entirety. Ultimately, a total of 9 full-text articles were considered for quantitative synthesis. This included 5 papers in English and 4 papers in Chinese. More specific information can be found in the study’s PRISMA flowchart (Figure 1).

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the selection process. RCT: randomized controlled trial.
Characteristics of Included Studies

We included 9 RCTs [34-42] from 6 countries. There were no statistical differences in general information between the trial and control groups in each study. A total of 791 study participants were included, with 395 in the experimental group taught CPR training using serious games and 396 in the control group taught CPR training using traditional methods. Additional information can be found in Multimedia Appendix 2 [34-42].

Quality Assessment

The risk of bias evaluation of the included literature is presented in Figure 2 [34-42] (the colors green, yellow, and red in the figure mean “low risk of bias,” “unclear risk of bias,” and “high risk of bias,” respectively). The quality of the included studies was found to be acceptable. In 6 RCTs [36-38,40-42], they described the generation of random sequences, of which 5 RCTs [36-38,40,41] described methods of allocation concealment. Due to CPR training and teaching, it was not possible to blind participants. In 3 RCTs [36-38], they applied the blinding method for researchers. Additionally, in 1 RCT [34], they had a high risk of reporting bias, and all 9 RCTs had complete data and did not have any other bias.

Meta-Analysis Results

The Effect of Serious Games Teaching on CPR Theory Performance

In the analysis, 6 out of the 9 studies [34,35,38,40-42] used posttraining CPR theory assessment as an outcome measure in RCTs. The pooled results revealed significant heterogeneity among the studies (P<.001; I²=93%), necessitating the use of a random-effects model for the meta-analysis. Figure 3 [34,35,38,40-42] demonstrates that there was no significant disparity in the theory assessment between the 2 groups under investigation (SMD –0.22, 95% CI –0.96 to 0.51; P=.55).

Figure 2. Methodological quality assessment of risk of bias for the included trials.

Figure 3. Meta-analysis of the effect of serious games on theory assessment.
The Effect of Serious Games Teaching on the Performance of CPR Skills Operations

Posttraining CPR skill manipulation performance was assessed as an outcome indicator in 5 RCTs [35,36,38,39,42] out of the 9 studies included. Meta-analysis was conducted using a random-effects model due to heterogeneity among the studies. Meta-analysis was conducted using a random-effects model due to heterogeneity among the studies ($P<.001; I^2=95\%$). The results indicated that there was no significant difference in skills assessment between the 2 study groups (SMD $-0.49$, 95% CI $-1.52$ to $0.55$; $P=.36$). This suggests that the use of serious games for CPR training did not lead to a significantly different skill level compared to other traditional training methods (Figure 4 [35,36,38,39,42]).

Figure 4. Meta-analysis of the effect of serious games on skill assessment.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Serious game</th>
<th>Traditional teaching</th>
<th>IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Sena 2019</td>
<td>8.4 0.21</td>
<td>23</td>
<td>22 15.6%</td>
</tr>
<tr>
<td>Hou 2015</td>
<td>67.4 7.6</td>
<td>35</td>
<td>63 21.1%</td>
</tr>
<tr>
<td>Huang 2021</td>
<td>60.24 5.29</td>
<td>50</td>
<td>85 23.1%</td>
</tr>
<tr>
<td>Wu 2016</td>
<td>94.65 6.02</td>
<td>65</td>
<td>99.4 21.4%</td>
</tr>
<tr>
<td>Yeung 2017</td>
<td>7.05 2.09</td>
<td>25</td>
<td>7.06 20.6%</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>198</td>
<td>194 $100.0%$</td>
<td>0.49 [1.52, 0.55]</td>
</tr>
</tbody>
</table>

Heterogeneity: $Q^2=1.26, Ch^2=0.27, df=4 (P<.0001); I^2=95\%$

Test for overall effect: $Z=0.92 (P=.36)$

The Effect of Serious Games Teaching on the Depth of CPR Compression

A total of 3 studies [36,37,39] presented findings on the impact of serious games on CPR compression depth. The assessment of heterogeneity demonstrated variability among the included studies ($P=.10; I^2=56\%$), necessitating the application of a random effects model. The analysis depicted in Figure 5 [36,37,39] revealed that the disparity between the 2 groups did not reach statistical significance (MD $3.17$, 95% CI $-0.18$ to $6.53$; $P=.06$).

Figure 5. Meta-analysis of the effect of serious games on cardiopulmonary resuscitation (CPR) compression depth.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Serious game</th>
<th>Traditional teaching</th>
<th>IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drummond 2017</td>
<td>41.42 13.04</td>
<td>40</td>
<td>43.71 15.29</td>
</tr>
<tr>
<td>Wu 2020</td>
<td>65.29 12.13</td>
<td>65</td>
<td>69.56 13.71</td>
</tr>
<tr>
<td>Yeung 2017</td>
<td>42.09 1.61</td>
<td>25</td>
<td>37.35 2.48</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>130</td>
<td>120 $100.0%$</td>
<td>3.17 [0.18, 6.53]</td>
</tr>
</tbody>
</table>

Heterogeneity: $Q^2=5.05, Ch^2=4.52, df=2 (P=.10); I^2=56\%$

Test for overall effect: $Z=1.85 (P=.06)$

The Effect of Serious Games Teaching on the Frequency of CPR Compression

A meta-analysis was performed on 3 studies [36,37,39] that investigated the impact of serious games training on the frequency of CPR compression. Due to the variation among these studies ($P=.005; I^2=81\%$), a random effects model was used. The results, as illustrated in Figure 6 [36,37,39], indicated that there was no significant difference in the theory of CPR compression rate between the 2 study groups (MD $-0.20$, 95% CI $-7.29$ to $6.89$; $P=.96$).

Figure 6. Meta-analysis of the effect of serious games on cardiopulmonary resuscitation (CPR) compression frequency.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Serious game</th>
<th>Traditional teaching</th>
<th>IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drummond 2017</td>
<td>102 19.09</td>
<td>40</td>
<td>111.94 28.46</td>
</tr>
<tr>
<td>Wu 2020</td>
<td>118.82 10.99</td>
<td>65</td>
<td>119.68 13.33</td>
</tr>
<tr>
<td>Yeung 2017</td>
<td>119.44 5.14</td>
<td>25</td>
<td>113.75 7.66</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>130</td>
<td>120 $100.0%$</td>
<td>-0.20 [-7.29, 6.69]</td>
</tr>
</tbody>
</table>

Heterogeneity: $Q^2=29.24, Ch^2=10.42, df=2 (P=.005); I^2=61\%$

Test for overall effect: $Z=0.06 (P=.96)$

Sensitivity Analysis

We conducted separate analyses using both fixed effects and random effects models to examine the SMD, MD, and 95% CI of each model. By systematically excluding studies one by one, when the study by de Sena et al [38] was excluded, we observed a decrease in heterogeneity from 93% to 0% for theoretical assessment (Figure 7 [34,35,40-42]) and from 95% to 54% for skill assessment (Figure 8 [35,36,39,42]), respectively. This indicates that the study conducted by de Sena et al [38] may have contributed to the observed heterogeneity. In the meta-analysis of CPR compression depth, heterogeneity decreased from 56% to 0% after the exclusion of the study by Drummond et al [37], indicating that this study was the source of heterogeneity (Figure 9 [36,39]). After the exclusion of the study by Yeung et al [36], the heterogeneity of the meta-analysis on compression frequency of CPR decreased from 81% to 56%, indicating that this study was one of the sources of heterogeneity (Figure 10 [37,39]).
Figure 7. Sensitivity analysis of meta-analysis of cardiopulmonary resuscitation (CPR) theory performance.

Figure 8. Sensitivity analysis of meta-analysis of cardiopulmonary resuscitation (CPR) skills operations.

Figure 9. Sensitivity analysis of meta-analysis of cardiopulmonary resuscitation (CPR) compression depth.

Figure 10. Sensitivity analysis of meta-analysis of cardiopulmonary resuscitation (CPR) compression frequency.

GRADE Evidence Quality Levels
Table S1 in Multimedia Appendix 3 presents the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) system evidence level for each outcome indicator in the meta-analysis of this study. The 4 outcome indicators considered were theory assessment, skill assessment, compression depth, and compression frequency.

Discussion
Principal Findings
This study systematically evaluated the efficacy of serious games-based training in CPR education, drawing upon data from 9 studies with a total of 791 participants. Our findings reveal no significant differences in theoretical exam scores, skill assessment scores, compression depth, or compression frequency between serious games-based and traditional CPR training methods. This suggests that serious games offer a highly effective alternative for CPR education. In alignment with the 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care [27], which recommend the incorporation of serious games into CPR education, our results underscore the positive impact of virtualized, gamified learning models on knowledge acquisition and CPR skill mastery. Theory and skills assessments are pivotal components of CPR training, serving as key indicators of training effectiveness and student proficiency. Our meta-analysis demonstrates that serious games-based CPR training is on par with traditional methods in enhancing both knowledge acquisition and skill levels. Consequently, serious games represent a valuable addition to the spectrum of CPR teaching and training methods, fostering innovation and aligning...
with the American Heart Association’s guidelines for modernizing teaching tools and approaches.

**Comparison With Previous Work**

This study aligns with a previous meta-analysis [43], indicating that both lay and medical school students exhibit enhanced knowledge following web-based digital resuscitation training. Moreover, they demonstrate comparable cognitive outcomes to those undergoing traditional training sessions. The inclination of younger individuals toward serious games for acquiring new skills stems from their immersive and interactive nature, offering a secure trial-and-error environment [44]. This, coupled with engaging and positive learning experiences, reduces reliance on educational resources and fosters active, independent learning—especially when conventional training methods are inaccessible. This approach helps in sustaining knowledge levels, preventing decay over time, and attaining learning outcomes equivalent to traditional education forms [45]. Despite these benefits, serious games’ applications for CPR training face challenges, presenting a mixed landscape concerning usability and enjoyment quality. Issues range from outdated guidelines and unupdated advice to overly detailed, professional information hindering learning efficacy. Such drawbacks may discourage public engagement with CPR learning [46]. For nonmedical learners, serious games must ensure acceptable usability, simplifying the comprehension and retention of CPR theoretical knowledge. Regular updates aligning with the latest guidelines can transform serious games into dynamic electronic textbooks [46,47]. To maximize the potential of serious games over traditional training, it is crucial to identify and evaluate functions that motivate learners to increase frequency and actively embrace knowledge updates. This strategic approach could position serious games as a superior alternative for enhancing the theoretical understanding of CPR, offering distinct advantages over traditional training methods [46,48].

Numerous guidelines [27,49-51] underscore the primary goal of CPR training: imparting participants with the skills necessary for high-quality CPR. This involves maintaining the correct compression rate and depth, ensuring thoracic recoil, and minimizing interruptions and hyperventilation. However, traditional training methods have presented challenges, particularly for nonmedical personnel [52,53], in mastering these vital competencies. Previous studies [52,54] have noted that simulated scenarios and repetitive practice often fall short of achieving adequate compression depth and frequency. Aksoy [55] and Siqueira et al [56] propose that a CPR teaching mode based on serious games could enhance learners’ motivation and attitude, consequently improving compression quality. This study echoes Lau et al’s [57] systematic review, indicating equivalence between serious games and traditional training methods in enhancing compression depth and frequency. However, electronic CPR training, including serious games, may not independently enhance skills without some influence from instructors, particularly for beginners. In other words, teacher involvement remains crucial to refining CPR skill performance through serious games training. Lim et al [25] discovered that content learned in serious games may not seamlessly transfer to skill operations during assessments, particularly for students with autonomous learning based on serious games. Scores in the pressing position, crucial for CPR quality, were notably worse than those in traditional training. Factors such as incorrect anatomical positions directly impact compression quality, making it challenging to achieve better performance in practical measures such as compression depth and frequency. While there was no significant difference in CPR compression skill or rate between the 2 training models, serious games-based CPR training revealed imperfections. To address this, integrating and emphasizing the impactful elements and advantageous attributes of traditional training into serious games may compensate for their shortcomings in skill practice. This approach has the potential to amplify the comparative advantages of serious games in CPR training.

In summary, the results of this study are similar to those of similar previous systematic reviews or studies. Nevertheless, given the limited number of studies included in this meta-analysis and the low GRADE evidence level, these results warrant cautious interpretation. Therefore, we recommend future CPR training efforts prioritize conducting high-quality, large-sample studies. This will enable a more comprehensive analysis of the effectiveness of serious games-based training, providing substantial evidence for the refinement of guidelines and the development of related teaching methodologies.

**Strengths and Limitations**

**Strengths**

This review compensates for the shortcomings of the previous literature in English by focusing on all types of serious games and conducting a comprehensive search of massive Chinese databases. Certainly, this study was conducted in strict accordance with highly recommended guidelines (ie, PRISMA), with early registration of the protocol for the systematic review and final grading of the evidence based on the GRADEpro GDT web-based tool, so it can be considered a robust, high-quality review. In addition, the meta-analysis conducted in this study involved 9 RCTs [34-42]. These RCTs provided detailed information on the study population, training protocol, serious games used, and measurement tools for outcome indicators. As for blinding implementation, it was challenging to blind interventionists due to the nature of CPR teaching training, which resulted in some degree of implementation bias. On the other hand, blinding the measurer effectively prevented measurement bias, particularly when assessing CPR theoretical knowledge and skills. Objective outcome indicators such as CPR compression depth and frequency, as recorded by the simulator, were less susceptible to measurement bias. The literature also addressed missed visits, had a low risk of selective reporting bias, and demonstrated baseline comparability between groups. Therefore, the included literature was of high quality, and the findings can be considered credible.

**Limitations**

This study acknowledges several limitations that merit consideration. First, our research only encompassed studies available in Chinese and English, which may introduce a linguistic bias. Second, heterogeneity in our meta-analysis results emerged due to variations in study populations, the use of different serious games, and diverse tools used to measure
outcome indicators. Despite our efforts to explore the sources of heterogeneity through sensitivity analysis, a complete explanation remained elusive. In particular, it is worth noting that the use of different instruments by the included studies to evaluate training outcomes may have influenced the judgment of the results. Third, the relatively small number of included studies prevented us from conducting tests for publication bias. Additionally, some data underwent statistical transformations during the meta-analysis, potentially influencing the accuracy of the results. Lastly, this study focused primarily on CPR theory assessment, skill evaluation, compression depth, and compression rate as outcome indicators, without delving into knowledge and skill retention post-training, trainees’ self-efficacy, or other facets of compression quality.

**Implications for Future Research and Practice**
Serious games, as an innovative model for CPR teaching and training, offer a promising avenue for first aid education, catering to diverse populations. However, this approach is still in its developmental and exploratory phases, and its cost-effectiveness warrants discussion. Future research should consider incorporating outcome indicators from the field of health economics to address economic barriers and promote the adoption of serious games in professional medical education and broader first aid training. Additionally, many studies lack standardized training specifications for serious games, including training duration, frequency, trainer intervention levels, and evaluation methods and tools for assessing training effectiveness. While serious games are recommended for CPR education, the specific details of this training mode require further standardization. Moreover, the quality of serious games, which serve as the platform for CPR training, significantly impacts training effectiveness. Developing serious games that align with international guidelines and cater to the diverse characteristics of trainees is undoubtedly challenging but essential. In conclusion, future research should prioritize conducting multicenter, large-sample RCTs to advance our understanding of the potential of serious games in CPR education.

**Conclusion**
This study conducted a meta-analysis of RCTs to assess the efficacy of serious games in CPR training. The findings indicate that serious games are equally effective as traditional training methods in enhancing CPR theory assessment and skill evaluation. Meanwhile, no significant differences emerged between serious games and traditional training methods regarding CPR compression depth and frequency. Notably, the current body of high-quality studies on serious games in CPR training is limited, often characterized by small sample sizes. Therefore, future research should prioritize conducting additional high-quality RCTs to provide further evidence and offer a more comprehensive understanding of the impact of serious games in CPR training and education.

**Acknowledgments**
This study was supported by the Hainan Provincial Natural Science Foundation of China (822MS071), the National Natural Science Foundation of China (81960420), and the Zhejiang Provincial Medical and Health Technology Project (2024KY1065).

**Data Availability**
Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

**Authors’ Contributions**
PC conceived the idea for the review. PC, YH, and PY conducted the data curation, methodology, validation, and formal analysis and wrote the first draft of the manuscript. PC, YH, PY, HW, and BX were involved in the study selection, quality assessment, and data extraction. PC, BX, and CQ conducted the statistical analysis. HZ is responsible for the writing, methodology, conceptualization, supervision, and editing of this manuscript.

**Conflicts of Interest**
None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 19 KB - games_v12i1e52990_app1.docx ]

Multimedia Appendix 2
Detailed characteristics of the included trials.
[DOCX File, 22 KB - games_v12i1e52990_app2.docx ]

Multimedia Appendix 3
Supplementary table S1.
[DOCX File, 15 KB - games_v12i1e52990_app3.docx ]
References


Aksoy ME, Ozkan AE, Kitapcioglu D, Usseli T. Comparing the outcomes of virtual reality-based serious gaming and lecture-based training for advanced life support training: randomized controlled trial. JMIR Serious Games 2023;11:e46964 [FREE Full text] [doi: 10.2196/46964] [Medline: 37768719]


Aksoy ME. Comparing basic life support serious gaming scores with hands-on training platform performance scores: pilot simulation study for basic life support training. JMIR Serious Games 2020;8(4):e24166 [FREE Full text] [doi: 10.2196/24166] [Medline: 33237035]


Review

Electronic Interactive Games for Glycemic Control in Individuals With Diabetes: Systematic Review and Meta-Analysis

WenQi Yao1*, BM; YiBing Han1*, MM; Li Yang1,2*, MD; Ying Chen1, MM; ShengZhe Yan1, BM; YanZhen Cheng1, MD

1Department of Endocrinology, Zhujiang Hospital of Southern Medical University, Guangzhou, China
2Department of Nutrition, Zhujiang Hospital of Southern Medical University, Guangzhou, China
*these authors contributed equally

Corresponding Author:
YanZhen Cheng, MD
Department of Endocrinology
Zhujiang Hospital of Southern Medical University
No.253 Gongye Avenue Central
Guangzhou, 510280
China
Phone: 86 0 206 278 2330
Email: chengyzx@163.com

Abstract

Background: Several electronic interventions have been used to improve glycemic control in patients with diabetes. Electronic interactive games specific to physical activity are available, but it is unclear if these are effective at improving glycemic control in patients with diabetes.

Objective: This study aimed to determine the effects of electronic game–based interventions on glycemic control in patients with diabetes.

Methods: Relevant studies that were published before April 1, 2023, were searched from 5 databases: PubMed, Embase, Web of Science, Scopus, and Cochrane Library. Eligibility criteria included prospective studies examining the relationship between electronic games with physical activities or diet education and glycemic control as the outcome. The risk of bias was assessed using the Cochrane risk-of-bias tool. All analyses were conducted using RevMan5.4.1. Depending on the heterogeneity across studies, the pooled effects were calculated using fixed-effects or random-effects models.

Results: Participants from 9 studies were included and assessed. Glycated hemoglobin (HbA1c) and fasting blood glucose improved in the intervention group, although the analysis revealed no significant reduction in HbA1c (−0.09%, 95% CI −0.29% to 0.10%) or fasting blood glucose (−0.94 mg/dL, 95% CI −9.34 to 7.46 mg/dL). However, the physical activity of individuals in the intervention group was significantly higher than that of those in the control group (standardized mean difference=0.84, 95% CI 0.30 to 1.38; P=.002). Other outcomes, such as weight and blood lipids, exhibited no significant improvement (all P>.05).

Conclusions: Electronic games had a good impact on participants’ physical activity and offered an advantage in glycemic control without reaching statistical significance. Electronic games are convenient for reminders and education. Low-intensity exercise games may not be considered a better adjuvant intervention to improve diabetes self-management care.

(JMIR Serious Games 2024;12:e43574) doi:10.2196/43574

KEYWORDS

electronic game; physical activity; diet; diabetes mellitus; glycemic control

Introduction

Diabetes mellitus is one of the 4 major noncommunicable diseases and is also among the top 10 global causes of death. Throughout the world, the number of patients with diabetes mellitus is increasing, probably due to changes in lifestyle. According to the International Diabetes Federation, in 2021, approximately 536.6 million adults (aged 20-79 years) were living with diabetes; this is expected to rise to 12.2% in 2045 [1]. Because of the rise in type 1 and type 2 diabetes, the burden of health care expenditures and its complications continues to increase, whereas the complications are the main causes of
morbidity and mortality [2]. To address the health challenge resulting from diabetes, effective and efficient management is needed [3-5].

Lifestyle management, an efficacious method for diabetes prevention [6], is a fundamental aspect of diabetes care. It includes diabetes self-management education and support, medical nutrition therapy, physical activity, smoking cessation counseling, and psychosocial care [7]. Food intake and physical activity are associated with significantly improved control of diabetes [8]. With advances in technology, lifestyle management incorporating novel technologies and formats meets the needs of various populations for diabetes treatment [9]. New methods, such as electronic games and wearable devices, aim to contribute to better patient compliance [10].

It has been reported that electronic games can help players learn more about healthy diets and encourage exercise [11,12]. Although they play a role as facilitators in motivating and accelerating physical activity, they offer little benefit to patients with chronic disease [13]. Previous systematic reviews have evaluated the impact of app-based or electronic health interventions to support changes in blood glucose management, physical activity, or diet [9,14,15]. However, previous papers analyzed relatively few articles or articles that were not solely on using games. They also used educational or regulation applications, robots, or virtual worlds that do not contain game elements. Electronic games specific to physical activity and dietary education are available; however, we currently lack an understanding of how effective electronic games can be for glycemic control.

In this study, we performed a comprehensive literature search to select studies on the effects of electronic game–based interventions on glycemic control in patients with diabetes for meta-analysis. Electronic gaming interventions are defined as containing an element of gaming that involves virtual reality, serious gaming, or exergaming [15].

**Methods**

**Data Sources and Search Strategy**

This review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement and its associated checklist (Multimedia Appendix 1). Relevant studies were published before April 1, 2023, were searched from 5 databases: PubMed, Embase, Web of Science, Scopus, and Cochrane Library. The references of the included studies were hand-searched to identify any additional articles. The following terms were used during the search: (“Diabetes” OR “diabetic” OR “diabetes mellitus” OR “glycemic control” OR “glucose control” OR “glucose”) AND (“game” OR “gamification” OR “exergaming” OR “avatar” OR “wii” OR “virtual” OR “konami” OR “wii-fit” OR “kinect” OR “tierone” OR “video-game” OR “serious-games” OR “serious video-games” OR “Augmented reality” OR “mixed reality” OR “second life” OR “TierOne” OR “Konami Dance Dance Revolution” OR “Sony Eyetoy” OR “Microsoft Kinec”). Detailed search strategies for each database are given in Multimedia Appendix 2. The reference lists of the searched articles and the relevant reviews were then screened to identify any pertinent studies.

**Study Selection**

Studies included in this meta-analysis met the following criteria: (1) participants were diagnosed as having type 1 diabetes or type 2 diabetes; (2) the articles were published in English or Chinese; (3) the articles presented the electronic management intervention with a gaming element, such as a virtual reality game, serious game, or exergame; and (4) the outcome indicators were blood glucose and glycated hemoglobin (HbA1c).

Studies that met the following criteria were excluded: (1) participants had gestational diabetes mellitus, had other special types of diabetes mellitus, underwent surgery, had an operation, or were in the emergency department; (2) participants had a previous history of mental illnesses, eating disorders, or cancer; (3) the management intervention was only based on an online, mobile, or virtual application but did not use a gaming element; and (4) articles that were protocols, conference abstracts, case reports, reviews, or meta-analyses.

Articles were screened in a 2-step process. First, all titles and abstracts were examined by 2 investigators. Any citations that clearly did not meet the inclusion criteria were excluded. Second, all abstracts and full-text articles were examined independently by 2 investigators. Any disagreements in the selection process were resolved through discussion with a third investigator.

**Risk of Bias**

The included trials were independently assessed by 2 investigators for the risk of bias using the Cochrane risk-of-bias tool [16]. An assessment was performed across 5 domains of bias (sequence generation, allocation concealment, blinding, incomplete outcome data, and selective reporting). The risk of bias was assessed as either low (proper methods taken to reduce bias), high (improper methods creating bias), or unclear (insufficient information provided to determine the bias level). All discrepancies and disagreements were resolved through consensus or, where necessary, by a third author.

**Data Extraction**

A Microsoft Excel table was used to extract data on the year of publication, country, sample size, participant characteristics, study setting and design, intervention and control arms, duration, and outcome data. The main outcomes included HbA1c or fasting blood glucose (FBG). The secondary outcomes included daily steps (regarded as a physical activity outcome), blood pressure, and weight, among others. The data were obtained from the original text and attachments supplied. Data from different studies were converted to common units. Data extraction was carried out by 2 reviewers independently. All discrepancies and disagreements were resolved through consensus.

**Missing Data**

Study authors were contacted by email where there were missing or unclear data (for instance, relating to the primary outcome). Studies for which insufficient primary data were available (eg,
missing data cannot be obtained) were excluded from analysis but not from the review.

**Data Synthesis and Quality Assessment**

All analyses were conducted using RevMan5.4.1 (Cochrane). Data were expressed as the mean difference (MD) and 95% CI and pooled using fixed-effect or random-effects models according to the heterogeneity. A random-effects model assumes that the study estimates are estimating different, yet related, intervention effects and thus incorporates the heterogeneity among studies. This is a more appropriate method to pool studies that may differ slightly in the distribution of risk factors, population, size, and outcomes.

Heterogeneity was assessed using a $\chi^2$ test and quantified using the $I^2$ statistic. Significance for the heterogeneity was set at $P<.05$, with an $I^2>50\%$ considered to be evidence of high heterogeneity, which prompted us to use the random-effects model to pool the data.

**Results**

**Overview**

Our search identified 10,088 articles, of which 4605 were screened after removing duplicate records. Of these, 182 were identified for further evaluation. Of these, 173 were excluded, resulting in 9 included studies (Figure 1). Of the excluded articles, 18 were excluded because they only had abstracts and we could not access the original text and data.

![Flowchart of the study selection process.](https://games.jmir.org/2024/1/e43574)

The results of the remaining 9 studies, comprising 913 participants and 747 cases of type 2 diabetes, were included in the meta-analysis [17-25]. The characteristics of all 9 studies are shown in Multimedia Appendix 3 [17-25]. The duration of trials ranged from 1 month to 1 year. Of the 9 studies, 4 were undertaken in the United States [18,21,23,25], 2 in Europe [17,20], and 3 in Asia [19,22,24]. Of the 9 studies, 3 assessed FBG and 8 assessed HbA$_1c$.

The studies included 2 non–randomized controlled trials (RCTs) and 7 RCTs, the quality of which was assessed using the Cochrane risk-of-bias tool. We determined that 3 studies were of high quality, whereas 4 were of moderate quality and 2 were of low quality (Figures 2 and 3 [17-25]). The 2 non-RCTs were not random and the allocations were unclear. Blinding was difficult in game interventions; 1 study was unblinded [25] and 4 were unclear, but the studies made an effort to blind either patients or personnel. One study was not blinded to the outcome assessment, but it was still analyzed as low risk, considering its main outcome was the objective index. The 9 studies had no elective outcome reporting.
Figure 2. Risk-of-bias graph showing the authors’ judgments about each risk-of-bias item presented as percentages across all the included studies. A total of 9 trials were assessed for risk of bias.

Figure 3. Risk-of-bias summary showing the authors’ judgments about each risk-of-bias item for each included study. Green “+”: low risk of bias; red “−”: high risk of bias; yellow “?”: unknown risk of bias.

Publication bias was not assessed for any outcome as <10 trials were available.

Meta-Analysis

HbA1c Level

A total of 8 articles had HbA1c testing but 1 did not provide postintervention data [25]. We sent an email to the author with a request to provide the raw data but received no reply.

As shown in Figure 4A [17,21,23-25], this analysis showed a clinically important improvement in HbA1c, but there was no significant reduction after the intervention among individuals with diabetes mellitus (7 studies; n=607; MD=−0.09%, 95% CI −0.29% to 0.10%; \( I^2 =37\% \); \( P=.36 \)). Figure 4B shows the change in HbA1c after a diet-based game intervention (3 studies; n=167; MD=−0.09%, 95% CI −0.48% to 0.30%; \( I^2 =2\% \); \( P=.65 \)). Figure 4C shows the change in HbA1c after a physical
activity–based game intervention (5 studies; n=508, $\text{MD}=-0.12\%$, $95\% \text{ CI} =-0.34\%$ to 0.09%; $I^2=51\%$; $P=.27$).

**Figure 4.** Meta-analysis of the effect of electronic games on HbA1c, FBG, and physical activity. (A) HbA1c after a diet intervention or physical activity intervention; (B) HbA1c after a diet intervention; (C) HbA1c after a physical activity game intervention; (D) FBG after an electronic game intervention; and (E) physical activity after an electronic game intervention. FBG: fasting blood glucose; HbA1c: glycated hemoglobin; IV: inverse variance; Std.: standardized.

**Fasting Blood Glucose Level**

The meta-analysis showed that the FBG level of the intervention groups was not statistically different from that of the control groups (3 studies; n=286; $\text{MD}=-0.94 \text{ mg/dL}$, $95\% \text{ CI} =-9.34$ to 7.46 $\text{mg/dL}$; $I^2=0\%$; $P=.83$; **Figure 4D**).

**Physical Activity**

Of the 7 RCTs, 2 assessed self-reported physical activity and 2 counted participants’ daily steps during the intervention to assess the patients’ physical activity. Because of the differences in measurement instruments, we calculated standardized mean differences (SMDs). These results were statistically heterogeneous with respect to the effect ($\chi^2=19.70$; $P<.001$; $I^2=85\%$); we found a significant increase in physical activity above baseline in the intervention groups. Moreover, participants assigned to the intervention groups increased their physical activity significantly more than participants in the control groups (SMD=0.84; $95\% \text{ CI } 0.30$ to 1.38; $P=.002$; **Figure 4E**).

**Weight**

Weight also trended toward decreases in the intervention groups, with an MD of $-1.46 \text{ kg}$ (95% CI $-4.71$ to 1.80 kg; **Figure 5A**).
However, the decreases did not reach statistical significance ($P=.38$).

**Figure 5.** Meta-analysis of the effect of electronic games on (A) weight, (B) total cholesterol, (C) LDL-C, (D) HDL-C, and (E) triglycerides. LDL-C: high-density lipoprotein cholesterol; IV: inverse variance; LDL-C: low-density lipoprotein cholesterol; Std.: standardized.

### Blood Lipids

There was no significant reduction in total cholesterol (3 studies; n=261; MD=0.05 mmol/L, 95% CI −0.22 to 0.33 mmol/L; $P=2%$; $I^2=76%$; $P=0.71$; **Figure 5B**), low-density lipoprotein cholesterol (4 studies; n=440; MD=0.08 mmol/L, 95% CI −0.09 to 0.24 mmol/L; $I^2=0%$; $P=36%$; **Figure 5C**), high-density lipoprotein cholesterol (3 studies; n=261; MD=0.02 mmol/L, 95% CI −0.06 to 0.10 mmol/L; $I^2=18%$; $P=61%$; **Figure 5D**), or triglycerides (3 studies; n=261; MD=0.02 mmol/L, 95% CI −0.32 to 0.37 mmol/L; $I^2=12%$; $P=89%$; **Figure 5E**) after the intervention among patients with diabetes mellitus.

### Discussion

This study demonstrated that electronic interactive games were associated with a good impact on participants’ physical activity. However, we found that electronic interactive games did not present a significant benefit for HbA1c levels, FBG levels, weight, or blood lipids compared to the control group. The game interventions were intended for education to manage diabetes through games.

### Effects of Diet Education Games on Blood Glucose

Plant-based diets and exercise are major diabetes-protective factors [26]. The Da Qing Diabetes Prevention Study showed an overall 51% reduction in diabetes incidence in participants after a 6-year intervention with diet, exercise, or both; its 30-year follow-up showed that lifestyle interventions reduced the incidence of serious diabetes complications and diabetes-related mortality [27]. However, Hemmingsen et al [28] did not find firm evidence that diet alone or physical activity alone influences the risk of type 2 diabetes mellitus or its associated complications in people at increased risk of developing type 2 diabetes mellitus compared to standard treatment [28]. The trials included in this study had little data on the impact of games on
diet, and only 3 articles evaluated participants' postintervention diet. From the results, education through games was effective, although the improvements in glycemic control were not statistically significant. The most important reason was that the 3 trials studied patients with type 1 diabetes mellitus aged 8 to 18 years. The games provided diabetes-related diet education to the patients, but family-based diet intervention may also not impact glycemic control [29].

Effects of Games Related to Physical Activity on Blood Glucose

Physical activity with different intensities impacts glycemic control in individuals with diabetes. Of the included studies, 4 trials [17,18,20,24] assessed physical activity by daily steps or self-reported activity, and this analysis found a significant increase. These results are consistent with findings from other meta-analyses showing increased physical activity among patients with chronic disease [30-32]. Some studies find positive effects with low-intensity physical activity, although these are not reflected by a decrease in HbA1c or FBG in patients with type 2 diabetes [33-35]. A meta-analysis showed that high-intensity interval exercise significantly reduced HbA1c levels compared to no or low-intensity exercise [36]. Low exercise intensity in the 9 studies we included may be the reason why there was no significant difference in HbA1c and FBG in patients with diabetes between the groups. However, the games in the virtual reality group were relatively novel, which was very helpful for improving cognition, physical skills that are directly involved in functional abilities, and enthusiasm for sports [19].

The study by Höchsmann et al [20] contributed a substantial amount of heterogeneity; without this study, I² was 11%. The high heterogeneity may have been caused by the baseline of the participants in this trial being better than those in the other trials. In their trial, Höchsmann et al [20] used a dilapidated garden to symbolize the patient’s physical condition, and exercise and daily physical activity execution were tracked by mobile phones, allowing for feedback. After 24 weeks of intervention, there was no significant change in HbA1c levels in the intervention group, while HbA1c levels in the control group receiving 1-time lifestyle counseling increased. In the trial, the intervention group had a higher increase in daily steps than the control group, providing evidence that physical activity can be encouraged by electronic games.

Effects of Games on Blood Lipids, Blood Pressure, and BMI

In our study, game-based intervention resulted in no significant decrease in blood lipids in patients with diabetes. Only 2 trials reported the outcomes of blood pressure [17,20] and BMI [17,19], and the 2 indexes were both reduced. Systolic blood pressure was below 140 mm Hg but above 130 mm Hg, which is still high for patients with diabetes. Treatment with medication may be indispensable.

Effects of Games on Weight

Lifestyle intervention can be effective for achieving clinically important reductions in body weight [37,38]. It has been demonstrated that electronic game activities are engaging, which encourages their use on a regular basis, improving the long-term outcome of a treatment for obesity [39,40]. However, an intervention using a different avatar did not improve physical activity practice or self-efficacy expectations [41]. Gomez et al [42] showed that high exercise intensity from active electronic games elicited significant increases in energy expenditure. In this study, electronic games did not result in significant weight reduction, and BMI was reduced slightly in 2 trials. Possible reasons include insufficient physical activity and that participants did not strictly control their diet. Whether electronic games are beneficial for weight control by encouraging appropriate intensity exercise in patients with diabetes requires more clinical evidence in the future.

The reasons for the lack of significant results in this meta-analysis may be as follows. First, participants in the control groups were also familiar with what the game taught. Second, participants with type 2 diabetes mellitus in the intervention groups, who were all older than 40 years, could not make full use of electronic devices and adapt to the games. Third, for exercise-based interventions, not all studies involved regular exercise monitoring for participants and established appropriate feedback or interaction mechanisms.

Limitations and Future Directions

This study had several limitations. First, not every included study reported the HbA1c and FBG levels. Some excluded studies had relevant interventions but did not observe blood glucose changes or failed to give detailed trial data results. Second, the studies that were included in this meta-analysis were not homogeneous. Different games or game mechanisms were used in different patient populations. The number of participants was not large in several of the included studies, and each study used different games. Therefore, it is difficult to conduct detailed hierarchical verification of the effects of different games on blood glucose. We strived to ensure that the included studies were high-quality RCTs with strict inclusion and exclusion criteria, excluding nongaming electronic interventions. Existing studies have evaluated the effectiveness of electronic games as an alternative for traditional diabetes education. As diabetes continues, it is necessary to promote this management model. However, future studies should not only design the game in terms of increased knowledge and improved self-management but should encourage enhanced physical activity intensity.

Conclusion

As an alternative treatment tool in diabetes management, the studies on electronic games explored in this study showed a clinical improvement in glycemic control and weight control, although this improvement was not superior to that observed in the control participants. Thus, such interventions may complement existing treatment courses for diet, self-management education, and high-intensity physical activity to potentially increase the compliance of patients with diabetes. More new technologies can be used for diabetes control, and electronic games can be designed for different groups of patients with diabetes. For example, immersive virtual reality is an emerging strategy to enhance exercise performance for young
patients with diabetes, and the metaverse may be a new community enabling older patients to form new social connections and share their experiences of living with diabetes. Interactive exercise games can be used in children to increase interest in education and family companionship time, and thus improve exercise compliance.

Acknowledgments
This work was supported by grants from the National Natural Science Foundation of China (grants 82170796 and 82003370), the Natural Science Foundation of Guangdong province (grant 2019A1515012225), the Science and Technology Plan of Guangdong province (grant 2022A0505030005), and Science and Technology Program of Guangzhou (grants 202102020165 and 2024A04J9990).

Authors' Contributions
Y Cheng conceived the idea of the overview and wrote the protocol and full overview. WY and YH developed the concept and details of the overview (ie, participants, intervention, comparison, outcomes). WY and Y Chen carried out searches and selected reviews for inclusion (YH acted as an arbitrator). WY and SY carried out the assessment of methodological quality (YH acted as an arbitrator). WY and SY extracted the data and interpreted the initial findings. YH and Y Chen directed data analyses. WY, YH, LY, and Y Cheng formulated the focus of the discussion and made suggestions for future studies. All authors were involved in the interpretation of the results and in approving the final review.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 28 KB - games_v12i1e43574_app1.docx ]

Multimedia Appendix 2
Search strategy.
[PDF File (Adobe PDF File), 116 KB - games_v12i1e43574_app2.pdf ]

Multimedia Appendix 3
Basic characteristics of the included studies.
[DOCX File , 21 KB - games_v12i1e43574_app3.docx ]

References


**Abbreviations**

- **FBG**: fasting blood glucose
- **HbA1c**: glycated hemoglobin
- **MD**: mean difference
- **PRISMA**: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
- **RCT**: randomized controlled trial
- **SMD**: standardized mean difference

©WenQi Yao, YiBing Han, Li Yang, Ying Chen, ShengZhe Yan, YanZhen Cheng. Originally published in JMIR Serious Games (https://games.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
Virtual Reality Therapy for the Management of Chronic Spinal Pain: Systematic Review and Meta-Analysis

Tongtong Zhang1,2*, BS; Xin Li1,2*, MS; Xuan Zhou1*, MM; Lixia Zhan3*, MM; Fan Wu1*, BM; Zefan Huang1, BM; Yuxun Sun4; Yufei Feng2; Qing Du1,5, PhD

1Department of Rehabilitation, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China
2School of Exercise and Health, Shanghai University of Sport, Shanghai, China
3The Second People’s Hospital of Beihai, Beihai, China
4College of Rehabilitation Sciences, Shanghai University of Medicine & Health Sciences, Shanghai, China
5Chongming Hospital, Shanghai University of Medicine & Health Sciences, Shanghai, China

* these authors contributed equally

Corresponding Author:
Qing Du, PhD
Department of Rehabilitation
Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine
1665 Kongjiang Road, Yangpu District, Shanghai, China
Shanghai, 200092
China
Phone: 86 021 25078600
Email: duqing@xinhuamed.com.cn

Abstract

Background: The effectiveness of virtual reality (VR) therapy in adults with chronic spinal pain (CSP) is unclear.

Objective: This study was conducted to compare the effectiveness of VR therapy and other therapies in adults with CSP, especially patients with inflammation-related pain.

Methods: PubMed, Web of Science, Cochrane Library, Embase, and CINAHL databases were searched up to November 11, 2023. Randomized controlled trials (RCTs) comparing adults with CSP receiving VR therapy with those receiving other therapies were included. The trial registration platform as well as the reference lists of included studies and previous systematic reviews and meta-analyses were manually searched. Two independent reviewers performed study selection, data extraction, risk-of-bias assessment, and evaluation of the quality of the evidence. The weighted mean difference (WMD) was used as the effect size used to synthesize the outcome measure.

Results: In total, 16 RCTs involving 800 participants were included in this meta-analysis. The pooled data from 15 (94%) RCTs including 776 (97%) participants showed that VR therapy was superior in improving pain intensity (WMD=–1.63, 95% CI –2.11 to –1.16, P<.001, I²=90%) and reducing inflammatory markers, including C-reactive protein (WMD=–0.89, 95% CI –1.07 to –0.70, P<.001, I²=90%), tumor necrosis factor-alpha (WMD=–6.60, 95% CI –8.56 to –4.64, P<.001, I²=98%), and interleukin-6 (WMD=–2.76, 95% CI –2.98 to –2.53, P<.001, I²=0%). However, no significant differences were found in terms of the spinal range of motion (ROM), disability level, or fear of movement. In addition, 10 (63%) of the included RCTs had a high risk of bias.

Conclusions: VR therapy may be an effective and safe intervention for reducing symptoms in patients with CSP, as it is shown to exert significant analgesic effects and beneficial improvements in inflammatory factor levels. However, this approach may not have significant effects on the spinal ROM, disability level, or fear of movement. Notably, the quality of the evidence from the RCTs included in this study ranged from moderate to low. Therefore, we recommend that readers interpret the results of this study with caution.

Trial Registration: PROSPERO CRD42022382331; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=382331

(JMIR Serious Games 2024;12:e50089) doi:10.2196/50089
KEYWORDS
virtual reality; chronic spinal pain; inflammation-related pain; systematic review; meta-analysis

Introduction
Chronic spinal pain (CSP), which most commonly includes chronic low back pain (CLBP) and chronic neck pain (CNP), is the leading cause of years with disability worldwide [1,2] and constitutes the most frequent reason for patients to seek medical care in any given year. The lifetime prevalence of low back pain (LBP) is 84%; more specifically, the lifetime prevalence of CLBP is 23%, and LBP accounts for approximately 11%-12% of cases of disability [3]. CSP is recognized as a biopsychosocial syndrome [4]. Prolonged pain can lead to anxiety, depression, and other negative emotions and is particularly significant in patients with CSP, as it is associated with decreased quality of sleep and reduced physical activity, thus placing tremendous strain on health care systems and world economies [5].

Previous studies have reported that an intervertebral disc undergoes aging or pathologic changes in the adjacent region in patients with CSP, exposing cells within the nucleus pulposus to macrophages, resulting in an inflammatory response that might trigger pain [6,7]. The guidelines recommend that nonsteroidal anti-inflammatory drugs (NSAIDs) be the primary choice for patients with chronic pain [8]. However, compared with a placebo, NSAIDs can reduce CSP by controlling the level of inflammation but do not achieve clinically important efficacy [9]. Additionally, long-term use may be associated with adverse effects (eg, gastrointestinal reactions, hepatic and renal damage, and cardiovascular risk) [10]. Several studies have shown that conventional nonpharmacological therapies, such as spinal manipulation, acupuncture, exercise therapy, yoga, and cognitive-behavioral therapy, are beneficial for reducing CSP and improving psychological symptoms but have limited effects (small to moderate) [11-14]. Effective cognitive-behavioral therapies are not widely accessible due to the reliance on therapist experience, and the long-term effectiveness of these therapies remains unclear [15]. Notably, the majority of patients with CSP have goals of pain management (using ongoing care) rather than “curing” (care with a specific end) for their therapeutic care because of the complexity of the causes of chronic pain [16]. Thus, pain management is as important as the control of inflammation levels for patients with CSP. There is an urgent need for an alternative analgesic nonpharmacological and anti-inflammatory strategy for patients with CSP.

Virtual reality (VR) is typically characterized by low cost, easy availability, reusability, and personalized customization; VR therapy has been used as an alternative approach for pain management in various populations, such as individuals with spinal cord injuries, burns, and phantom limb pain [17-19]. VR can be categorized into 2 types: nonimmersive virtual reality (NIVR) and immersive virtual reality (IVR). NIVR is managed using a computer or console gaming system and a 2D interface device (mouse, keyboard, or gamepad, joystick), and patients do not need to be fully immersed in a virtual environment for experience [20]. With the use of professional equipment, hardware, and configuration of the corresponding software, IVR can mimic reality by enabling the user to interact with the virtual environment [21]. A recent study demonstrated that regular exercise with the use of VR might be related to a decrease in inflammation in participants undergoing chronic hemodialysis [22], and inflammatory arthritis—targeting innovative teaching approaches based on VR technology are considered feasible [23]. There is limited evidence regarding the beneficial effects of VR therapy on pain in patients with CNP [24] and CLBP [25,26]; furthermore, there is insufficient focus on inflammatory factors. Therefore, this study aimed to investigate the potential efficacy of VR in reducing pain intensity and the levels of inflammatory factors in patients with CSP, thereby providing an updated summary of the existing evidence.

Methods

Study Protocol and Registration
This systematic review and meta-analysis was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines. The PRISMA checklist is given in Multimedia Appendix 1. The study protocol was registered in the PROSPERO database (CRD42022382331). The Cochrane Handbook for Systematic Reviews of Interventions (version 5.1.0) was followed [27].

Search Strategy

Search Sources
PubMed, Web of Science, Cochrane Library, Embase, and CINAHL electronic databases were searched from inception to November 11, 2023, to identify relevant studies. The reference lists of the included studies, as well as systematic reviews and meta-analyses that examined the efficacy of VR in patients with CSP, were manually searched for additional eligible studies. The trial registration platform ClinicalTrials was also searched for ongoing studies that reported sufficient data on the efficacy of VR for CSP.

Search Terms

The studies on VR for CSP were identified by formulating appropriate search terms. These terms were selected based on the target population (spinal pain, neck pain, thoracic pain, back pain, LBP, sacral pain, and intervertebral disc pain), target intervention (eg, VR), and target study design (eg, randomized controlled trial [RCT]). The detailed search strategy is shown in Multimedia Appendix 2.

Study Eligibility Criteria

The inclusion criteria were as follows:
- Participants: adults older than 18 years with chronic pain (more than 12 weeks) in the spinal region were included, except those who were receiving analgesic medication and who had cancer-related pain or neuropathic pain (eg, neuropathic pain after spinal cord injury, herniated disc with compression, sciatica, or lumbar sacral radiculitis).

https://games.jmir.org/2024/1/e50089 JMIR Serious Games 2024 | vol. 12 | e50089 | p.199 (page number not for citation purposes)
• Intervention: VR therapy.
• Comparisons: sham stimulation, usual care, and conventional treatment.
• Outcomes: pain intensity, inflammatory markers (e.g., C-reactive protein [CRP], tumor necrosis factor-alpha [TNF-α], and interleukin [IL]-2, IL-4, and IL-6), fear of movement, spinal range of motion (ROM), and disability level.
• Study design: RCT.

No restrictions were imposed on language or publication date.

Study Selection
The retrieved studies were imported into Endnote X9 software (Clarivate), which was used to eliminate duplicate studies. Two independent reviewers (authors TTZ and FW) performed the initial screening of the literature by reading the titles and abstracts of all retrieved studies, and studies that did not meet the inclusion criteria were excluded. Next, the full texts of the remaining studies were screened. Any disagreements were resolved by negotiation and discussion with a third reviewer (author XZ).

Data Extraction
Two independent reviewers (authors FW and XL) extracted detailed information, including the name of the first author, the year and country of publication, the language of publication, study design, the number of included subjects (% female), diagnosis, and outcome indicators. Information about the characteristics of the interventions, including dose, frequency, and duration, was also collected for both the VR group and the control group. The sample size and mean (SD) of the outcome indicators in each group were collected. When the same group of participants was reported in different studies, the group with the largest sample size was selected for inclusion in this review to avoid duplicate publications [28]. For information that could not be confirmed, the authors were contacted by email. The 2 reviewers cross-checked the data at the end of the extraction, and any disagreements were resolved by negotiation.

Risk-of-Bias Assessment
The methodological quality of the included studies was independently assessed by 2 reviewers (authors XL and ZFH) using the Cochrane Risk of Bias tool, and the studies were classified as having a low, unclear, or high risk of bias [29]. Disagreements were resolved by consulting a third reviewer (author QD). The Egger test and funnel plots generated with Stata 14.0 software (StataCorp) were used to evaluate potential publication bias. The trim-and-fill method was used to adjust for funnel plot asymmetry due to publication bias [30]. Sensitivity analyses were performed by removing each study separately to assess the robustness of the results [29]. The overall strength of the evidence was assessed using the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) criteria [31].

Meta-Analysis and Subgroup Analysis
This systematic review and meta-analysis were performed using Review Manager software 5.4 (Informer Technologies) and Stata 14.0 software. Heterogeneity was tested using the $I^2$ statistic. A fixed effects model was selected for the outcome indicators if $I^2<50\%$, while a random effects model was used when there was significant statistical heterogeneity ($I^2>50\%$, $P<.05$). The effect size used to synthesize the outcome measure was the weighted mean difference (WMD). Three subgroup analyses were performed to explore the possible causes of heterogeneity among the studies: the region of CSP (CNP vs CLBP), VR types (IVR vs NIVR), and treatment duration (<4 weeks vs ≥4 weeks).

Results

Search Results
A total of 924 records were obtained from the 5 databases and the trial registration platform. A total of 394 (42.6%) duplicates were identified and removed using Endnote X9 software. After screening the titles and abstracts, 40 (7.5%) of the remaining 530 RCTs were retained, and 490 (92.5%) were excluded for the following reasons: (1) the study population included patients without CSP, (2) the intervention did not use VR therapy, (3) the type of study was a non-RCT, (4) the information was incomplete, and (5) the patients also received analgesic medication. Of the 40 studies, 15 (38%) were retained after reading the full text and 25 (62%) were excluded for the following reasons: (1) the study population included patients without CSP, (2) the intervention did not use VR therapy, (3) the type of study was a non-RCT, (4) the information was incomplete, and (5) the patients also received analgesic medication. Two additional RCTs were retrieved from the reference lists of the included studies. One RCT was retained after the full text was read, and the other was excluded due to incomplete information. A total of 16 studies were included in this review, 15 (94%) of which reported sufficient data (e.g., mean [SD], sample size) on the analgesic effect of VR for CSP. Therefore, 15 studies were included in the meta-analysis. The PRISMA flowchart of selecting the included studies is shown in Figure 1.
The CSP reported in the included studies included CLBP [32-43] and CNP [44-47]. All patients had chronic pain that persisted for more than 3 months. The sample size varied from 8 to 90 participants, and the mean age ranged from 18 to 85 years. The characteristics of all the studies are summarized in Table 1.
Table 1. Characteristics of the included studies [32-47].

<table>
<thead>
<tr>
<th>First author</th>
<th>Patient characteristics</th>
<th>Diagnosis</th>
<th>Outcome measures</th>
<th>Time points</th>
<th>Dropout rate</th>
<th>Country, language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garcia et al [32]</td>
<td>T*: 179 I*: 89 (75) C*: 90 (78)</td>
<td>CLBP&lt;sup&gt;d&lt;/sup&gt;</td>
<td>DVPRS&lt;sup&gt;e&lt;/sup&gt;, Pain Catastrophizing Scale (PCS), 8-item Chronic Pain Acceptance Questionnaire (CPAQ-8)</td>
<td>Baseline, –7, 0, 4, 7, 11, 14, 18, 21, 25, 28, 32, 35, 39, 42, 46, 49, 53, 56 days</td>
<td>I: 0 C: 0</td>
<td>United States, English</td>
</tr>
<tr>
<td>Nambi et al [33]</td>
<td>T: 60 I (VR&lt;sup&gt;f&lt;/sup&gt;): 20 I (core stabilization [CS]): 20 C: 20</td>
<td>CLBP</td>
<td>NPRS&lt;sup&gt;g&lt;/sup&gt;, quality of life (physical fitness index)</td>
<td>Baseline, 4 weeks, 8 weeks, 6 months</td>
<td>I (VR): 0.05 I (CS): 0.05</td>
<td>Saudi Arabia, English</td>
</tr>
<tr>
<td>Nambi et al [34]</td>
<td>T: 45 I (VR): 15 I (isokinetic training [IKT]): 15 C: 15</td>
<td>CLBP</td>
<td>NPRS</td>
<td>Baseline, 4 weeks</td>
<td>I (VR): 0 I (IKT): 0</td>
<td>Saudi Arabia, English</td>
</tr>
<tr>
<td>Yalfani et al [35]</td>
<td>T: 25 I: 13 C: 12</td>
<td>CLBP</td>
<td>VAS&lt;sup&gt;h&lt;/sup&gt;, 36-item Short Form Health Survey (SF-36)</td>
<td>Baseline, 8 weeks</td>
<td>I: 0 C: 0</td>
<td>Iran, English</td>
</tr>
<tr>
<td>Park et al [36]</td>
<td>T: 24 I (NWE): 8 I (lumbar stabilization exercise [LSE]): 8 C: 8</td>
<td>CLBP</td>
<td>VAS</td>
<td>Baseline, 8 weeks</td>
<td>I: 0 C: 0</td>
<td>South Korea, English</td>
</tr>
<tr>
<td>Afzal et al [37]</td>
<td>T: 90 I: 45 (64.28) C: 45 (69.04)</td>
<td>CLBP</td>
<td>VAS, Modified Oswestry Disability Index</td>
<td>Baseline, 4th, 8th 12th sessions</td>
<td>I: 0.07 C: 0.07</td>
<td>Pakistan, English</td>
</tr>
<tr>
<td>Nambi et al [38]</td>
<td>T: 60 I (VRE): 20 I (isokinetic exercise [IKE]): 20 C: 20</td>
<td>CLBP</td>
<td>VAS, inflammatory biomarkers</td>
<td>Baseline, 4 weeks</td>
<td>I (VRE): 5 I (IKE): 5</td>
<td>Saudi Arabia, English</td>
</tr>
<tr>
<td>Nambi et al [39]</td>
<td>T: 36 I (VR): 12 I (combined physical rehabilitation [CPR]): 12 C: 12</td>
<td>CLBP</td>
<td>Inflammatory biomarkers</td>
<td>Baseline, 4 weeks</td>
<td>I (VR): 0 I (CPR): 0</td>
<td>Saudi Arabia, English</td>
</tr>
<tr>
<td>Nambi et al [40]</td>
<td>T: 54 I (VR): 18 I (CPR): 18 C: 18</td>
<td>CLBP</td>
<td>VAS, TSK-17</td>
<td>Baseline, 4 weeks</td>
<td>I (VR): 0 I (CPR): 0</td>
<td>Saudi Arabia, English</td>
</tr>
<tr>
<td>First author</td>
<td>Patient characteristics</td>
<td>Outcome measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matheve et al [41]</td>
<td>T: 84</td>
<td>NPRS, Roland-Morris Disability Questionnaire (RMDQ), PCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 42 (64)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 42 (64)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age (years), mean (SD)</td>
<td>CLBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 42.1 (11.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 44.2 (11.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline, postintervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dropout rate (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stamm et al [42]</td>
<td>T: 22</td>
<td>NRS&lt;sup&gt;g&lt;/sup&gt;, Chronic Pain Grade Questionnaire (CPGQ), 12-item Short Form Health Survey (SF-12), Hannover Functional Ability Questionnaire for Measuring Back Pain–Related Disability (Fib-H-R), TSK&lt;sup&gt;l&lt;/sup&gt;-11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 11 (73)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 11 (55)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>CLBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 75.00 (5.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 75.50 (4.39)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline, 4 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dropout rate (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monteiro-Junior et al [43]</td>
<td>T: 34</td>
<td>NRS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 17 (100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 17 (100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>CLBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 68 (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cetin et al [44]</td>
<td>T: 41</td>
<td>NRS&lt;sup&gt;l&lt;/sup&gt;, Joint position sense error (JPSE), VAS, pressure pain threshold (PPT), SF-36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>CNP&lt;sup&gt;m&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 40.00 (11.88)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 41.94 (10.76)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline, 6 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dropout rate (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahat et al [45]</td>
<td>T: 90</td>
<td>NDI&lt;sup&gt;j&lt;/sup&gt;, VAS, EQ-5D, TSK-17, cervical range of motion (CROM), kinematic measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (VR): 30 (63)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (laser): 30 (70)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>CNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (VR): 48.00 (14.07)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (laser): 48.00 (17.41)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 48.00 (17.76)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline, 4 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dropout rate (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (VR): 16.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (laser): 13.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 16.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nusser et al [46]</td>
<td>T: 55</td>
<td>NRS, active cervical range of motion (ACROM), NDI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (VR): 17 (53)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (sensorimotor group [SM]): 16 (69)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (SM): 53.1 (5.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 49.8 (8.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>CNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 51.2 (8.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 53.1 (5.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline, 3 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dropout rate (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (VR): 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I (SM): 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tejera et al [47]</td>
<td>T: 44</td>
<td>VAS, conditioned pain modulation (PPT), ACROM device, NDI, PCS, 11-item Spanish version of the TSK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 22 (50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 22 (54.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>CNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 32.72 (11.63)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 26.68 (9.21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline, 4 weeks, 1 month, 3 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dropout rate (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I: 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C: 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>T: total participants.
<sup>b</sup>I: intervention group.
<sup>c</sup>C: control group.
<sup>d</sup>CLBP: chronic low back pain.
<sup>e</sup>DVPRS: Defense and Veterans Pain Rating Scale.
<sup>f</sup>VR: virtual reality.
<sup>g</sup>NPRS: Numerical Pain Rating Scale.
<sup>h</sup>VAS: Visual Analogue Scale.
<sup>i</sup>VRE: virtual reality exercise.
<sup>j</sup>NDI: Neck Disability Index.
The types of VR interventions included IVR [32,35,42,44-47] and NIVR [33,34,36-41,43], which were classified based on the degree of isolation participants experienced when interacting with the virtual environment during VR therapy. NIVR uses a wall-mounted screen or a computer monitor as the vehicle for VR content, while IVR uses a headset or head-mounted display [48]. Compared to NIVR, IVR can increase the user’s sense of presence by improving immersion through the addition of auditory or haptic feedback [49]. The duration of a single VR session ranged from 2 to 40 minutes, and the frequency of treatment ranged from 5 to 7 times a week; all the included studies ranged in duration from a single exercise session to 8 weeks. For the control groups, 5 (31%) studies performed conventional balance function training [33,34,38-40], 5 (31%) performed conventional physical therapy [36,37,41,46,47], 2 (13%) performed core training [43,44], and the remaining conducted treatments, including sham VR [32], conventional multimodal pain therapy [42], waiting lists [45], and standard care [35]. The intervention details are summarized in Table 2.
<table>
<thead>
<tr>
<th>First author</th>
<th>Intervention group</th>
<th>Control group</th>
<th>Device</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garcia et al [32]</td>
<td>Ease VR, IVR</td>
<td>Sham VR, NIVR, not interactive, displayed 2D nature footage with neutral music, 20 videos rotated over 56 sessions, performed 56 times (2-16 minutes each time, average of 6 minutes, 1 time/day)</td>
<td>Pico G2 4K all-in-one head-mounted VR device</td>
<td>8 weeks</td>
</tr>
<tr>
<td>Nambi et al [33]</td>
<td>VR group: sit in the virtual platform and select firing game executed by trunk movements (flexion, extension, and lateral flexion; 30 minutes/day, 5 times/week, for 4 weeks); heat modality (20 minutes); therapeutic ultrasound (25 minutes)</td>
<td>Conventional balance function training, traditional active balance exercise for abdominal and back muscles (5 times/week for 4 weeks); heat modality (20 minutes); therapeutic ultrasound (25 minutes)</td>
<td>VR group: Pro-Kin system PK 252N (TecnoBody)</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Nambi et al [34]</td>
<td>VRT: shooting game (30 minutes, 5 days/week, for 4 weeks); home-based exercise; hot-pack therapy (20 minutes); ultrasound (frequency 1 MHz, intensity 1.5 W/cm² in continuous form for 5 minutes)</td>
<td>Conventional balance function training: standardized conventional exercises actively involving abdominal, deep abdominal, and back muscles (30 minutes/session, 5 days/week, for 4 weeks); hydrocollator packs (20 minutes/session); continuous ultrasound (frequency 1 MHz, intensity 1.5 W/cm²) at the low back region (5 minutes, 5 days/week, for 4 weeks)</td>
<td>VRT: Pro-Kin system (TecnoBody)</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Yalfani et al [35]</td>
<td>Fishing, boxing, tennis, football, bowling, beat saber, audio shield, and skiing (30 minutes, 3 times/week, for 8 weeks)</td>
<td>Standard care.</td>
<td>VR: HTC Vive virtual reality system</td>
<td>8 weeks</td>
</tr>
<tr>
<td>Park et al [36]</td>
<td>NWE; using the Nintendo Wii exercise program, including the wakeboard, Frisbee dog, jet ski, and canoe games. Participants chose which Nintendo Wii sports program to perform and took a 2-minute break every 10 minutes (30 minutes/session, 3 times/week, for 8 weeks)</td>
<td>Conventional physical therapy: using physical agent modalities, such as a hot pack (30 minutes); interventional current therapy (15 minutes); deep heat with ultrasound (5 minutes)</td>
<td>VR: Nintendo</td>
<td>8 weeks</td>
</tr>
<tr>
<td>Afzal et al [37]</td>
<td>Kinetic exergames (trunk slide flexion, sitting to avoid obstacles, jumping and combined movement of arms, for 5 minutes); after 30 seconds of rest, play body ball game for 5 minutes (3 sessions/week for a total of 12 sessions); routine physical therapy</td>
<td>Conventional physical therapy: heat therapy for 10 minutes, hamstring stretching, back-strengthening exercises (3 sessions/week for a total of 12 sessions)</td>
<td>VR: nonimmersive system with a kinetic device (model V2), incorporated with red-green-blue (RGB) cameras and time-of-flight (TOF) sensor, attached with a liquid crystal display (LCD) screen</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Nambi et al [38]</td>
<td>VRE; virtual training exercises performed in the upright position, a car race game chosen from the list of games, and training given to focus on the back muscles. The participant was asked to sit on the moving game chair and instructed to watch the game on the desktop monitor (30 minutes/session, 5 days/week, for 4 weeks); hydrocollator packs (20 minutes/session); continuous ultrasound (frequency 1 MHz, intensity 1.5 W/cm²) at the low back region (5 minutes, 5 days/week, for 4 weeks)</td>
<td>Conventional balance function training: standardized conventional exercises actively involving abdominal, deep abdominal, and back muscles (30 minutes/session, 5 days/week, for 4 weeks); hydrocollator packs (20 minutes/session); continuous ultrasound (frequency 1 MHz, intensity 1.5 W/cm²) at the low back region (5 minutes, 5 days/week, for 4 weeks)</td>
<td>VRE: Pro-Kin system (TecnoBody)</td>
<td>4 weeks</td>
</tr>
<tr>
<td>First author</td>
<td>Intervention group</td>
<td>Control group</td>
<td>Device</td>
<td>Duration</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Nambi et al [39]</td>
<td>Virtual reality training (VRT): shooting game, sitting on a virtual platform and visualizing the game on the computer display screen (30 minutes each time, 5 times/week, for 4 weeks); heat modality (20 minutes); therapeutic ultrasound (frequency 1 MHz, intensity 1.5 W/cm²; 5 minutes); home-based exercise (10 repetitions, bottom-to-heel stretch, opposite arm/leg raise, back extension, bridging, knee rolling; 2 times/day for 4 weeks)</td>
<td>Conventional balance function training: active isotonic and isometric exercises for abdominal, deep abdominal, and back muscles (10-15 repetitions/day, 5 days/week for 4 weeks; stretching focused on each muscle group for 3 repetitions for 10 seconds per muscle group); heat modality (20 minutes); therapeutic ultrasound (frequency 1 MHz, intensity 1.5 W/cm²; 5 minutes); home-based exercise (10 repetitions, bottom-to-heel stretch, opposite arm/leg raise, back extension, bridging, knee rolling; 2 times/day for 4 weeks)</td>
<td>VR: Pro-Kin system PK 252 N (Pelvic Module balance trunk MF; TecnoBody)</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Nambi et al [40]</td>
<td>VRT: shooting game, sitting on a virtual platform and visualizing the game on the computer display screen (30 minutes each time, 5 times/week, for 4 weeks); heat modality (20 minutes); therapeutic ultrasound (frequency 1 MHz, intensity 1.5 W/cm²; 5 minutes)</td>
<td>Conventional balance function training: active isotonic and isometric exercises for abdominal, deep abdominal, and back muscles (10-15 repetitions/day, 5 days/week for 4 weeks; stretching focused on each muscle group for 3 repetitions for 10 seconds per muscle group); heat modality (20 minutes); therapeutic ultrasound (frequency 1 MHz, intensity 1.5 W/cm²; 5 minutes)</td>
<td>VR: Pro-Kin system PK 252 N (Pelvic Module balance trunk MF; TecnoBody)</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Matheve et al [41]</td>
<td>2 different games (2 minutes each); single-session intervention, 2 × 2 minutes of pelvic tilt exercises in the sagittal plane, with 30 seconds of rest in between, through a wireless motion sensor</td>
<td>Conventional physical therapy: 2 different games (2 minutes each); single-session intervention, 2 × 2 minutes of pelvic tilt exercises in the sagittal plane, with 30 seconds of rest in between</td>
<td>VR: wireless motion sensor (Valedo Pro, Hocoma)</td>
<td>Single exercise session</td>
</tr>
<tr>
<td>Stamm et al [42]</td>
<td>Multimodal pain therapy in VR (movement therapy and psychoeducation), training session including 12 exercises, structured as follows: (1) warm-up (training of upper and lower extremities), (2) main part (strengthening of abdominal and back muscles, core stability), (3) cool-down (stretching, progressive muscle relaxation), (4) psychoeducative units (topics: physiology of pain, pain management, stress management, everyday training), 3 times/week for 30 minutes</td>
<td>Conventional multimodal pain therapy: chair-based group exercises and psychoeducation in a group setting), 3 times/week for 30 minutes</td>
<td>VR: head-mounted display headset using the VIRST VR app</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Monteiro-Junior et al [43]</td>
<td>Virtual physical training (8 exercises, 30 minutes each time, with 3 weekly sessions lasting 90 minutes each, lasted 8 weeks, 3 times weekly/session</td>
<td>Core training: postures adopted by participants for 15-30 seconds or according to the capacity of each; 10-15 seconds between postures (ie, bridges), with each performed 3 times, lasted 8 weeks, 3 times weekly/session</td>
<td>VR: Wii Balance Board (WBB; Nintendo)</td>
<td>8 weeks</td>
</tr>
<tr>
<td>Cetin et al [44]</td>
<td>VR exercises: VR apps that allowed neck movements in all directions, motor control (MC) exercises (20 minutes and then VR for 20 minutes, 5 repetitions for each exercise; 40 minutes/session, 3 sessions/week, for 6 weeks, total of 18 sessions)</td>
<td>Core training: strengthening of deep cervical flexors (DCF), deep cervical extensors (DC), and axioscapular muscles; stretching exercises; and postural correction exercises (40 minutes, 10 repetitions for each exercise, 3 sessions/week, for 6 weeks, total of 18 sessions)</td>
<td>VR: Oculus Go VR glasses, 2 VR apps installed: “Ocean Rift” and “Gala 360”</td>
<td>6 weeks</td>
</tr>
<tr>
<td>First author</td>
<td>Intervention group</td>
<td>Control group</td>
<td>Device</td>
<td>Duration</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Bahat et al [45]</td>
<td>VR group: kinematic home training and customized software with the virtual airplane controlled by head motion (5 minutes, 4 times/day, 20 minutes/day, 4 times/week, for 4 weeks)</td>
<td>Waiting list</td>
<td>VR: customized neck VR system (hardware including Oculus Rift DK1 head-mounted display equipped with 3D motion tracking; software developed using Unity-pro, version 3.5, Unity Technologies)</td>
<td>4 weeks</td>
</tr>
<tr>
<td>Nusser et al [46]</td>
<td>VR group: neck-specific sensorimotor training (NSST)—head-repositioning test (HRT), head-to-target test (HTT), dynamic exercise including 5 different trajectories (3 minutes given between tasks), training divided into 6 20-minute sessions for a total of 120 minutes); standard rehabilitation program</td>
<td>Conventional physical therapy: different forms of general and neck-specific exercise therapies (strengthening, mobilization, relaxation, medical training therapy, functional gymnastics, aqua therapy, physical therapy, and traditional “back school”)</td>
<td>VR: modified VR system (Fraunhofer Institute für Graphische Datenverarbeitung), helmet (Schutz helm uvex pheos alpine, Fürth), 3Space Fastrak System (Pothemus Inc)</td>
<td>3 weeks</td>
</tr>
<tr>
<td>Tejera et al [47]</td>
<td>VR mobile apps “Full Dive VR,” only lateral flexion movements of the neck; “VR Ocean Aquarium 3D”: flexion, extension, and rotation movements (3 series of 10 repetitions, with 30 seconds of rest between exercises)</td>
<td>Conventional physical therapy: flexion, extension, rotation, and tilt exercises (3 series of 10 repetitions, with 30 seconds of rest between exercises)</td>
<td>VR: VR Vox Play glasses with a head-mounted display clamping system (weight 330 g) with an LG Q6 smartphone attached, 2 VR mobile apps installed</td>
<td>4 weeks</td>
</tr>
</tbody>
</table>

aVR: virtual reality.  
bIVR: immersive virtual reality.  
cNIVR: nonimmersive virtual reality.  
dNWE: Nintendo Wii exercise.  
eVRE: virtual reality exercise.

The risk of bias in the 16 (100%) studies included in the meta-analysis is presented in Figure 2. Overall, 10 (63%) studies showed a high risk of bias. In addition, 15 (94%) RCTs generated an adequately randomized sequence, and 9 (60%) of them were analyzed using a blinded method for outcome measurement. Ratings using the GRADE methodology for all outcome measurements were inconsistent and ranged from moderate to low quality (Multimedia Appendix 3). Therefore, the quality of evidence from most studies was classified as fair.
Primary Outcome

Pain Intensity

All 16 (100%) studies (800 patients) reported pain intensity: 9 (56%) used the Visual Analogue Scale (VAS) [34-38,40,44,45,47], 2 (13%) used the Numerical Pain Rating Scale (NPRS) [33,41], 3 (19%) used the Numeric Rating Scale (NRS) [42,43,46], and 1 (6%) used the Defense and Veterans Pain Rating Scale (DVPRS) [32]. The random effects model revealed that compared with the control treatment, the VR intervention significantly reduced pain intensity (WMD=–1.63, 95% CI –2.11 to –1.16, P<.001, I²=90%). Clinical differences between groups were significant, and as suggested, the minimal clinically important difference (MCID) threshold on the VAS for LBP was set at a 1.5-point reduction [50]. Given the significant heterogeneity observed (I²=90%), we performed subgroup analyses to investigate the source of heterogeneity based on the different regions, VR types, and treatment durations.

VR had a good analgesic effect on both CNP and CLBP groups compared with the control group. The results did not significantly differ among the subgroups (WMD=–1.63, 95% CI –2.11 to –1.16); see Figure 3. Moreover, a total of 7 (44%) studies demonstrated that IVR significantly improved CSP (WMD=–1.50, 95% CI –2.45 to –0.55, P<.001, I²=80%) [32,35,42,44-47]. Another 8 (50%) studies showed that NIVR improved CSP substantially (WMD=–1.50, 95% CI –2.45 to –0.55, P<.001, I²=90%) [33,34,36-38,40,41,43]; see Figure 4. The subgroup analyses also revealed significant differences between treatment durations of <4 weeks (WMD=–1.41, 95% CI –2.12 to –0.69, P<.001, I²=0%) and ≥4 weeks (WMD=–1.65, 95% CI –2.16 to –1.14, P<.001, I²=91%) in terms of the analgesic effect of VR treatment on CSP (Multimedia Appendix 4).
Secondary Outcomes

**Inflammatory Markers**

Patients with CSP develop a systemic inflammatory response and have elevated levels of inflammatory markers in the blood [51]. Two studies (62 patients) focused on the levels of inflammatory markers (eg, CRP, TNF-α, IL-2, IL-4, and IL-6) by collecting 10 mL of venous blood [38,39]. The results showed that VR therapy significantly improved the level of CRP (WMD=–0.89, 95% CI –1.07 to –0.70, P<.001, I²=0%).
TNF-α (WMD=–6.60, 95% CI –8.56 to –4.64, \(P<.001, \hat{I}^2=98\%\)), and IL-6 (WMD=–2.76, 95% CI 2.98 to –2.53, \(P<.001, \hat{I}^2=0\%\)). No significant differences were found between the IL-2 and IL-4 subgroups (Figure S1 in Multimedia Appendix 5).

**Fear of Movement**

Four studies (162 patients) reported fear of movement according to the 11-item or 17-item Tampa Scale of Kinesiophobia (TSK-11 or TSK-17, respectively) [42,47]. No significant differences were found in either the TSK-11 (WMD=–0.81, 95% CI –4.48 to 2.86, \(P=.66, \hat{I}^2=0\%\); Figure S2 in Multimedia Appendix 5) or TSK-17 (WMD=–9.66, 95% CI –22.01 to 2.68, \(P=.13, \hat{I}^2=97\%\); Figure S3 in Multimedia Appendix 5).

**Spinal Range of Motion**

Three studies reported changes in the ROM of the neck in 4 directions before and after the intervention [45-47]. No significant differences were found between the groups in terms of flexion (WMD=2.67, 95% CI –2.31 to 7.64, \(P=.29, \hat{I}^2=61\%\)), extension (WMD=3.92, 95% CI –2.17 to 10.0, \(P=.21, \hat{I}^2=48\%\)), right rotation (WMD=–0.22, 95% CI –4.38 to 3.95, \(P=.92, \hat{I}^2=0\%\)), or left rotation (WMD=0.08, 95% CI –3.90 to 4.05, \(P=.97, \hat{I}^2=42\%\)); see Figure S4 in Multimedia Appendix 5.

**Disability Level**

Three studies (139 patients) reported disability levels in patients with CNP by using the Neck Disability Index (NDI) [45-47], a 10-item questionnaire that assesses self-reported disability related to CNP. Higher scores on the NDI indicate higher levels of disability. No significant differences were found in the pooled analysis of 3 (19%) studies (WMD=–2.66, 95% CI –5.47 to 0.15, \(P=.06, \hat{I}^2=48\%\)); see Figure S5 in Multimedia Appendix 5.

**Adverse Events**

One study reported that after 1 month of intervention, patients experienced nausea and motion sickness [32], two studies reported that there were no adverse events [33,37], and the remaining studies did not mention adverse events. The overall dropout rate was 4.25% (17/400) in the intervention group and 3.75% (15/400) in the control group.

**Publication Bias and Sensitivity Analysis**

The Egger test indicated significant publication bias in the results for pain intensity (\(P=.03\); Figure 5). The sensitivity analysis for pain intensity revealed that removing each study separately did not significantly affect the pooled results, thus indicating that the results are robust (Figure 6). The trim-and-fill method was performed, and it was estimated that there were 4 missing studies. The pooled estimates (95% CIs) calculated for the fixed effects model and the random effects model were –2.30 (–2.42 to –2.18) and –2.06 (–2.50 to –1.61), respectively (Figure 7). No significant changes in the results were observed before or after pruning or filling, indicating that our results are robust and plausible.

**Figure 5.** Funnel plot of pain intensity in the VR group compared with the control group. VR: virtual reality; WMD: weighted mean difference.
**Discussion**

**Principal Findings**

The primary purpose of this meta-analysis was to compare the relative efficacy of VR therapy and other therapies (e.g., conventional therapy, sham stimulation, and standard care) for treating CSP. The results indicated that VR therapy can effectively relieve CSP. The results of subgroup analyses showed that VR is a beneficial pain management strategy for patients with CNP and CLBP. For different types of VR, subgroup analyses showed that compared to the control group, IVR and NIVR both significantly improved CSP. No statistically

---

**Figure 6.** Sensitivity analysis of the included studies.

**Figure 7.** Trim-and-fill analysis to estimate the number of potentially missing studies about the effect of VR on CSP. Circles represent real studies, and squares represent studies estimated by this method. Circles represent real studies, and squares represent studies estimated by this method. CSP: chronic spinal pain; VR: virtual reality.
significant differences were found between patients who underwent VR treatments for a duration of ≤4 weeks and a duration of ≥4 weeks. VR was associated with a significant improvement in inflammatory marker levels but not in the fear of movement, spinal ROM, or disability level. VR was found to be well tolerated among these patients.

Discussion of the Results

The primary result suggested that VR reduces self-reported pain intensity, which might be explained by several implicating mechanisms [52,53]. A previous study reported that abdominal muscle strength is significantly lower in people with LBP [54], and a lack of strength in the core trunk muscles can lead to a decrease in intra-abdominal pressure, affecting spinal stability [55]. VR, as a novel human-computer interaction approach, can stimulate and mobilize the sensory system during training and results in changes in neuroplasticity and enhanced performance of relevant muscle groups, promoting a new motor learning process and leading to increased spinal stability [37,56], which would benefit pain relief. Furthermore, previous studies have reported that an intervertebral disc undergoes aging or pathological changes in the adjacent region in patients with CSP, exposing cells within the nucleus pulposus to macrophages, resulting in an inflammatory response that might trigger pain [7,8]. VR therapy may enhance the activity of disc fibroblasts and increase the thickness of the multifidus muscle [39,57], which is beneficial for relieving pain intensity. Furthermore, pain is an unpleasant subjective sensation associated with actual or potential tissue damage and is correlated with the degree of patient attention given to the pain area [58-61]. The various virtual game environments and real-time feedback methods are the most eye-catching features in the VR training process; these methods can be used to attract the patient’s visual and auditory attention to achieve motor performance, while relatively less attention has been given to the effects of VR on pain [62,63].

Although the high heterogeneity of the primary outcome and the results of the subsequent subgroup analyses suggest that the region of CSP, VR type, and treatment duration may play a role in the heterogeneity, the results of the sensitivity analysis indicate that these differences are more likely to be caused by 6 studies [33-35,37,38,40], which included participants of different ages.

VR therapy significantly improved the levels of inflammatory markers, including CRP, TNF-α, and IL-6. Numerous studies have previously reported an association between CSP and changes in inflammatory cytokines, such as IL-1 and TNF-α, which are thought to be closely related to the pathogenesis of disc herniation and degeneration [64,65]. Similarly, Nambi et al [66] reported that 4 weeks of VR training could significantly decrease pain intensity, increase functional impairment, and improve CRP, TNF-α, IL-2, IL-4, and IL-6 levels. However, the limited number and low quality of the included studies need to be noted, and further RCTs with large samples and rigorous study designs are needed to elucidate these results.

Patients with CSP may engage in fear/avoidance behaviors to avoid pain and protect themselves by limiting spinal motion, which ultimately affects spinal mobility and the speed of movement [67,68], with the degree of pain catastrophizing being proportional to the degree of disability [69,70]. However, we found no statistically significant differences in fear avoidance beliefs after the VR intervention but at the 3-month follow-up [47]. A systematic review and meta-analysis reported that VR therapy enhances spinal ROM and physical functioning in patients with CNP [26]. We failed to observe significant differences in the spinal ROM or disability level after VR intervention compared to those in the control group, which may be attributed to the relatively short duration (0-8 weeks) of the VR intervention (the reported mean duration was 4.81 weeks).

Limitations

Several limitations need to be addressed in this meta-analysis. First, the pooled analysis of the studies may be imprecise due to the large heterogeneity and the low quality of evidence from most of the included studies, and the results should be interpreted with caution. Second, the optimal duration of treatment for CSP could not be determined. Third, the effectiveness of VR therapy in patients with CSP and its analgesic effects in long-term follow-up must be further explored in high-quality studies. Fourth, indicators related to quality of life, such as depression and anxiety, should be emphasized and investigated in depth in future studies of patients with CSP.

Conclusion

VR therapy is an innovative and effective analgesic method that has beneficial effects on inflammatory markers in patients with CSP compared to other therapies (sham stimulation, usual care, conventional treatment). However, this approach may not have significant effects on the fear of movement, spinal ROM, or disability level. Notably, the quality of the evidence from the RCTs included in this study ranged from moderate to low. Therefore, we recommend that readers interpret the results of this study with caution. Future trials with large sample sizes, rigorous designs, and long-term follow-up periods are needed to explore the clinical significance of these differences and key issues in patients with CSP and to elucidate the underlying mechanisms of VR.

Acknowledgments

The authors thank Zhengquan Chen for providing support and advice. This work was supported by the Shanghai Hospital Development Center Foundation (SHDC12023118), the Medical Innovation Research of Shanghai Science and Technology Commission (22Y21900600), and the Three-year Action Plan of Shanghai Municipality to Further Accelerate the Inheritance, Innovation and Development of Traditional Chinese Medicine (2021-2023; ZY (2021- 2023)- 0201–05).
Data Availability
All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions
All authors contributed to the writing and redrafting of the manuscript. QD and XZ had the original idea. TZ and FW performed the literature search, XL and ZH assessed the risk of bias; YS, YF, and LZ rated the certainty of the evidence for each outcome; and XL and FW undertook data collection. The results were analyzed, interpreted, and discussed by XZ and QD. All authors contributed to the conception and design of the study, the analysis and interpretation of data, and the drafting and revising of the manuscript and have approved the final version.

Conflicts of Interest
None declared.

Multimedia Appendix 1
The PRISMA checklist. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis.

Multimedia Appendix 2
Search strategy for all electronic databases.

Multimedia Appendix 3
The GRADE criteria. GRADE: Grading of Recommendations, Assessment, Development and Evaluation.

Multimedia Appendix 4
Forest plots of the effect of VR compared with other treatments for pain intensity in patients with CSP: subgroup analysis of posttreatment effectiveness for treatment duration. CSP: chronic spinal pain; VR: virtual reality.

Multimedia Appendix 5
Forest plots of the effect of VR compared with other treatments for inflammatory marker level, fear of movement, spinal ROM, and disability level in patients with CSP. CSP: chronic spinal pain; ROM: range of motion; VR: virtual reality.

References


Abbreviations

CLBP: chronic low back pain
CNP: chronic neck pain
CRP: C-reactive protein
CSP: chronic spinal pain
DVPRS: Defense and Veterans Pain Rating Scale
GRADE: Grading of Recommendations, Assessment, Development and Evaluation
IL: interleukin
IVR: immersive virtual reality
LBP: low back pain
NDI: Neck Disability Index
NIVR: nonimmersive virtual reality
NPRS: Numerical Pain Rating Scale
NRS: Numeric Rating Scale
NSAID: nonsteroidal anti-inflammatory drug
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
RCT: randomized controlled trial
ROM: range of motion
TNF-α: tumor necrosis factor-alpha
TSK: Tampa Scale of Kinesiophobia
VAS: Visual Analogue Scale
VR: virtual reality
VRE: virtual reality exercise
WMD: weighted mean difference
Abstract

Background: The increasing prevalence of autism spectrum disorder (ASD) has driven research interest on the therapy of individuals with autism, especially children, as early diagnosis and appropriate treatment can lead to improvement in the condition. With the widespread availability of virtual reality, augmented reality (AR), and mixed reality technologies to the public and the increasing popularity of mobile devices, the interest in the use of applications and technologies to provide support for the therapy of children with autism is growing.

Objective: This study aims to describe the literature on the potential of virtual reality, AR, and mixed reality technologies in the context of therapy for children with ASD. We propose to investigate and analyze the temporal distribution of relevant papers, identify the target audience for studies related to extended reality apps in ASD therapy, examine the technologies used in the development of these apps, assess the skills targeted for improvement in primary studies, explore the purposes of the proposed solutions, and summarize the results obtained from their application.

Methods: For the systematic literature review, 6 research questions were defined in the first phase, after which 5 international databases (Web of Science, Scopus, ScienceDirect, IEEE Xplore Digital Library, and ACM Digital Library) were searched using specific search strings. Results were centralized, filtered, and processed applying eligibility criteria and using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The results were refined using a technical and IT-oriented approach. The quality criteria assessed whether the research addressed ASDs, focused on children’s therapy, involved targeted technologies, deployed solutions on mobile devices, and produced results relevant to our study.

Results: In the first step, 179 publications were identified in Zotero reference manager software (Corporation for Digital Scholarship). After excluding articles that did not meet the eligibility or quality assessment criteria, 28 publications were finalized. The analysis revealed an increase in publications related to apps for children with autism starting in 2015 and peaking in 2019. Most studies (22/28, 79%) focused on mobile AR solutions for Android devices, which were developed using the Unity 3D platform and the Vuforia engine. Although 68% (19/28) of these apps were tested with children, 32% (9/28) were tested exclusively by developers. More than half (15/28, 54%) of the studies used interviews as an evaluation method, yielding mostly favorable although preliminary results, indicating the need for more extensive testing.

Conclusions: The findings reported in the studies highlight the fact that these technologies are appropriate for the therapy of children with ASD. Several studies showed a distinct trend toward the use of AR technology as an educational tool for people with ASD. This trend entails multidisciplinary cooperation and an integrated research approach, with an emphasis on comprehensive empirical evaluations and technology ethics.
KEYWORDS

autism; autistic; autism spectrum disorder; ASD; virtual reality; augmented reality; extended reality; mixed reality; mobile app; children; preschool; mobile phone

Introduction

Background

In recent years, there has been increased interest in using technology to address the unique challenges faced by individuals with autism spectrum disorder (ASD). Among the various technological approaches, extended reality (XR), which includes augmented reality (AR) and virtual reality (VR), has emerged as a promising solution for intervention and therapy in children with ASD. XR offers the potential to create immersive and engaging environments that can address the specific needs of individuals on the autism spectrum, assisting them with communication, social interaction, and skill development. As a result, researchers and practitioners have explored the development of XR-based mobile apps tailored to the therapy of children with ASD.

However, the rapid growth in this field has spawned a multitude of XR-based mobile app solutions, each claiming unique benefits and features. With this proliferation of interventions, it is important to comprehensively assess the current landscape of XR-based mobile apps for the therapy of children with ASD, not only to strengthen existing knowledge in the field but also to provide critical insights into the research field.

In light of these considerations, this systematic literature review aimed to explore and assess the current status of XR-based mobile app solutions for the therapy of children with ASD. By synthesizing evidence from existing studies, this review aimed to provide an updated overview of the field, identify research gaps, and provide valuable insights. In this endeavor, this review aimed to contribute to the advancement of knowledge and practice in the field of XR-based interventions for ASD therapy, ultimately aiming to improve the quality of life of children on the autism spectrum and their caregivers.

ASD is a neurological condition that has a significant negative impact on a person’s social, verbal, and physical abilities. Researchers claim that ASD is typically discovered around the third year of life [1], but it can be identified and diagnosed as early as the age of 18 months [2]. According to a study from 2022, a total of 1% of infants have ASD [3]. On the basis of studies conducted over the past 50 years, the World Health Organization predicts a global increase in the prevalence of ASD [3].

Researchers consider 3 techniques that could be used to facilitate the evaluation of this condition’s prevalence: providing diagnostic tools, enhancing diagnostic standards, and increasing public awareness of ASDs [4].

On the basis of each individual’s verbal IQ and level of language delay, the diagnosis assesses the severity of the disorder (mild, moderate, or severe) [5], which can estimate the extent to which daily life is affected. Many people with ASD are timid, have difficulty communicating, or experience anxiety when engaging in casual conversation. Despite their communication and social skill deficits, individuals with ASD have demonstrated a preference for technology [6]. Furthermore, the use of technology in behavioral therapy for people with ASD has the additional advantage of being cost-effective in terms of both caregiver and treatment facility expenses [7]. In addition, various studies have shown that individuals with ASD respond better to visual stimuli than to other sensory stimuli [8]. The findings of these studies have led to various applications of technology in digital behavioral treatment.

XR, AR, VR, and Mixed Reality

As evidenced, the evolution of IT has accelerated in recent years. According to the study by Abad-Segura et al [9], rapid technological advancements have caused a significant and positive shift in how people view modern living.

The relatively new term XR refers to the entire spectrum from AR to VR, including mixed reality (MR; Figure 1 [10]).

Figure 1. The extended reality concept [10].

Extended reality

AR is a technology that enables real-time interaction and integration of 3D virtual models into the physical world [11]. Although the first portable AR system was developed in 2003 [12], AR did not acquire widespread acceptance and public...
awareness until the release of the mobile game Pokémon GO [13] in 2016. Despite the various implementation challenges, AR has many potential applications. In addition to applications in specific domains such as industry, construction, or medicine, as well as in advertising and commerce, education, and gaming [14], AR can be used by a broader audience for everyday tasks such as finding information about nearby points of interest, navigation, and assistance while following a route [14]. Most of these apps are now accessible owing to advances in mobile device technology and the spread of smart mobile phones. AR facilitates behavioral therapy by enhancing the experiences and abilities of people with ASD and establishing an integrated learning environment that enables the visualization of educational materials in 3D and engaging manipulation of real-world objects [15]. By generating “physical” structures to improve specific skills, AR fosters the imagination of patients with ASD without impairing it [16,17]. Moreover, AR can be used to create more engaging and appealing user interfaces, thereby eliminating the need for conventional input devices such as a keyboard and mouse [18]. AR technology is typically accessed using various devices and platforms. Among the widely used platforms and tools for developing AR apps are Unity, Unreal Engine, ARCoRe, and HP Reveal.

As described in the study by Azuma [11], VR is a computer-generated environment that simulates real-life scenarios, creating an immersive and interactive experience. This means that the users are placed in a completely virtual world, which can be similar to or different from the real one. This technology requires specialized equipment, such as VR headsets or glasses to enable users to see and interact with the virtual environment.

As seen in the studies by Bursali and Yilmaz [19] and El-Jarn and Southern [10], MR is situated between AR and VR, integrating the 2 technologies to provide the user with a unique and captivating experience in real time. It can be difficult to precisely define the limits of MR as they depend on the devices and equipment used as well as the extent to which VR or AR is incorporated into the final product. A model describing the integration of digital objects from the physical world into the virtual world is shown in the study by Milgram and Kishino [20], which also presented a taxonomy for MR, stating that it can be defined as a part of the human-computer interface field, which integrates VR and AR elements to create an environment in which virtual and real objects coexist and interact.

To be used, technologies from the XR spectrum require specific hardware with an optical sensor [19]. In addition, well-known technology companies such as Google, Facebook, Apple, Amazon, and Microsoft have significantly contributed to the development of AR tools and services [21,22], including handheld devices; holographic screens (Microsoft HoloLens); and heads-up displays, which are mainly designed for MR, tablets, and mobile devices (smartphones).

The development of collaborative XR, which enables simultaneous communication and collaboration among multiple users, is one of the research trends in the field of XR [23].

Given the benefits that AR, VR, and MR can offer as a new mode of human-computer interaction and the fact that these technologies are becoming ubiquitous and part of our daily lives, this systematic review aimed to describe how these technologies can be used in the therapy of children with ASDs.

Methods

Overview

According to Kitchenham [24], a systematic literature review is a method for identifying, evaluating, and interpreting all available research relevant to a field of study as well as answering specific research questions (RQs). We conducted this research following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [25] and the recommendations suggested by Kitchenham [24].

To conduct this literature review, several well-known scientific databases were queried, and publications containing relevant information for our analysis were filtered. We defined RQs and provided answers to each of them, thus achieving our proposed objective.

Search Strategy

According to the considered methodology, the following 7 RQs were formulated. These questions consider aspects relevant to the understanding of concepts important to this study:

1. What is the papers’ distribution over time? (RQ 1)
2. What category of people are the studies aimed at? (RQ 2)
3. Which technologies are used with XR or any of its subdivisions to develop apps for ASD therapy? (RQ 3)
4. What skills were targeted for improvement in primary studies? (RQ 4)
5. What are the purposes for which the proposed solutions were used? (RQ 5)
6. What are the results obtained using the proposed solutions? (RQ 6)

To study the literature and answer the aforementioned questions, we searched for scientific publications using various academic research databases. Our study primarily focused on the technical aspects of mobile app solutions using XR for autism therapy. To comprehensively cover our research domain, we chose to use multidisciplinary scientific databases—Scopus, ScienceDirect, and Web of Science—along with 2 databases particularly relevant to computer science, namely, IEEE Xplore Digital Library and ACM Digital Library. From these sources, we only considered publications that were relevant in computer science–related categories, such as technology, engineering, and computer science, excluding categories related to medicine, chemistry, or neurosciences considering that the RQs were focused not only on the available apps but also on their technical details. The functionalities, the technologies used, and the entire process of their development also constituted an objective. Thus, the approach from a technical point of view and the development of these apps were followed. This was done using the results refinement interface available in the aforementioned databases. Initially, to view and analyze the results of queries conducted using the considered search strings, the search was not limited to a particular time.
Given the topic of this study, we aimed to query scientific databases so that the resulting list of publications would meet the following criteria:

1. Reference to ASDs
2. Consideration of one of the technologies that are part of the concept of XR (VR, AR, or MR)
3. Addressing mobile apps
4. Aim to develop solutions for the therapy of children

The literature was searched using keywords relevant to achieving the proposed objectives: autism, autistic, ASD, virtual reality, augmented reality, extended reality, mixed reality, mobile application, and children.

Following the analysis of these keywords, the query process was extended by including the following terms: Autis*, VR, AR, MR, XR, Mobile app*, Smartphone app*, Child*, Infan*, Toddler*, Preschool*, Kid*, and Juvenile. In the aforementioned list, an asterisk stands for any number of characters at the end of the current string (eg, Preschool* refers to Preschool, Preschooler, and Preschoolers).

### Information Sources

Depending on the search options available in each database considered, specific search strings were defined for querying the databases (Table 1). These query strings were defined using advanced search functions and appropriate operators. The search of Web of Science and Scopus publications was performed by title, abstract, and keywords, and IEEE Xplore Digital Library, ScienceDirect, and ACM Digital Library were searched using a general search. The queries were executed on December 18, 2022.

#### Table 1. The search strings used for querying the databases (N=219).

<table>
<thead>
<tr>
<th>Item</th>
<th>Database</th>
<th>Search string</th>
<th>Returned results, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Web of Science</td>
<td>(TS=(Autis*) OR TS=(ASD)) AND (TS=(virtual reality OR VR) OR TS=(augmented reality OR AR) OR TS=(mixed reality OR MR) OR TS=(extended reality OR XR) AND (TS=(mobile app* OR smartphone app*)) AND (TS=(child* OR infan* OR toddler* OR preschool* OR kid* OR juvenile))</td>
<td>45 (20.5)</td>
</tr>
<tr>
<td>2</td>
<td>Scopus</td>
<td>(TITLE-ABS-KEY (Autis*) OR TITLE-ABS-KEY (asd)) AND (TITLE-ABS-KEY (“virtual reality” OR vr) OR TITLE-ABS-KEY (“augmented reality” OR ar) OR TITLE-ABS-KEY (“mixed reality” OR mr) OR TITLE-ABS-KEY (“extended reality” OR xr)) AND (TITLE-ABS-KEY (“mobile app*” OR “smartphone app*”) AND (TITLE-ABS-KEY (child* OR infan* OR toddler* OR preschool* OR kid* OR juvenile))</td>
<td>32 (14.6)</td>
</tr>
<tr>
<td>3</td>
<td>IEEE Xplore</td>
<td>(Autis* OR ASD) AND (Augmented reality OR AR OR Mixed reality OR MR OR Extended Reality OR XR OR Virtual Reality OR VR) AND (Mobile OR Tablet OR Smartphone OR Phone OR Smartglass) AND (App* OR Solution*) AND (child* OR kid* OR infan* OR preschool* OR juvenile OR toddler*)</td>
<td>26 (11.9)</td>
</tr>
<tr>
<td>4</td>
<td>ScienceDirect</td>
<td>(“Autism Spectrum Disorder” OR “ASD”) AND (“Augmented reality” OR “Mixed reality” OR “Extended reality”) AND (“App OR Application”) AND (“kids OR children”)</td>
<td>48 (21.9)</td>
</tr>
</tbody>
</table>

#### Eligibility Criteria

The papers obtained by querying scientific databases had an interdisciplinary nature. However, our study took a technical and IT-focused approach to mobile app solutions using XR for autism therapy. Therefore, we needed to refine the results by considering inclusion and exclusion criteria. As previously stated, no constraints were imposed on the publication dates of the articles during the search conducted in the scientific databases. Nevertheless, considering the significant progress in mobile device capabilities and their widespread use over the last decade, which have facilitated the development and growth of the global use of XR-based mobile apps for therapeutic purposes, we focused our investigation on the period following 2012 [26,27]. In line with our technical focus on mobile app solutions using XR for the therapy of children with ASD, we refined the search results across the 5 considered databases, prioritizing computer science-related domains. We deliberately excluded categories related to medicine, chemistry, or neurosciences as our RQs focused on both the available apps and their technical details.

Before centralizing the results for analysis, they were refined according to the inclusion and exclusion criteria. The inclusion criteria were as follows:

1. Articles published in English
2. Articles published after 2012

The exclusion criteria were as follows:

1. Book chapters
2. Paper tables of contents
3. Articles published in languages other than English
4. Results on the topics of medicine, chemistry, or neurosciences

After initial processing, the database searches returned the number of results presented in Table 2.
Table 2. Results obtained after initial data processing (N=179).

<table>
<thead>
<tr>
<th>Database</th>
<th>Results, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web of Science</td>
<td>43 (24)</td>
</tr>
<tr>
<td>Scopus</td>
<td>22 (12.3)</td>
</tr>
<tr>
<td>IEEE Xplore Digital Library</td>
<td>25 (14)</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td>41 (22.9)</td>
</tr>
<tr>
<td>ACM Digital Library</td>
<td>48 (26.8)</td>
</tr>
</tbody>
</table>

Selection Process

During the selection process, the PRISMA guidelines were considered [25]. These guidelines outline 3 steps: identification (centralizing the results and excluding duplicate publications), screening (review of titles and abstracts and testing eligibility), and inclusion (the publications identified as answering the proposed RQs). The PRISMA 2020 checklist is available in Multimedia Appendix 1. The Zotero reference manager software (Corporation for Digital Scholarship) was used to perform the specified steps. In the first step, 179 publications resulting from the search of the 5 considered databases were imported, after which duplicate publications (n=20, 11.2%) and some conference papers (n=4, 2.2%) were removed. For the next step, 86.6% (155/179) of the publications were considered. To enable author collaboration, the data were imported into Google Sheets. In total, 2 reviewers (M-VT and CET) conducted an independent screening of publications for inclusion based on title and abstract analysis. Studies meeting the eligibility criteria according to both reviewers were then considered for full-text screening. Any disagreements were discussed face-to-face between the reviewers, and a third party was involved to help reach unanimity where necessary. The same process was implemented for the full-text review with the assistance of a third reviewer (SV).

After reviewing the titles and abstracts, a total of 41.9% (65/155) of the publications were excluded as they did not address the proposed topic, focusing either on another condition or on other technologies.

Despite the high quality of the publications, as evidenced by their indexing in prestigious international databases, the analysis included a full text review (where available) of the remaining 58.1% (90/155) of the publications, and the following quality assessment criteria were applied to ensure their relevance to the RQs considered. Articles with no full text accessible were excluded. The following quality criteria (QCs) were applied to 78 publications:

1. Does the research topic address ASDs? (QC 1)
2. Does the study address children’s therapy? (QC 2)
3. Does the study include one of the technologies targeted in this review? (QC 3)
4. Is the solution deployed on a mobile device? (QC 4)
5. Are the results relevant to this review? (QC 5)

Given the nature of the research and its objectives, the 5 QCs that were developed specifically to achieve the goals of the research were used to evaluate the studies’ quality by 2 authors. Each publication was carefully reviewed and assigned a score from 0 to 2 measuring the extent to which it corresponded to the quality assessment criteria and to the subject of this study (0=no; 1=partially; 2=yes). Thus, the maximum score for a paper could be 10. After that, the data were combined for further analysis using Microsoft Excel (Microsoft Corp). After consolidation, a score with a decimal part (eg, 1.5) was rounded up to the nearest integer for better inclusion.

After reviewing the full texts, 3 publications were found to be not written in English, and in other publications, ASD was not addressed (only mentioned), other technologies were addressed, or the proposed solution was not clearly described and the results were inconclusive (see the sample in Figure 2; the entire table is available in Multimedia Appendix 2 [28-104]).

Following this, only publications with a score of ≥7 were evaluated as they adequately addressed the RQs. The type of publication, whether it was a review or aimed at developing an app, was also noted. Systematic literature review publications were investigated to identify any references that could be added to this study, but they were removed from the list after being reviewed. The entire publication selection process is illustrated in Figure 3.
Figure 2. Screenshot of the quality assessment of the papers.
Synthesis Methods

The data extraction process was carried out to methodically address the stated RQs. Initially, a single reviewer handled the task of data extraction, leveraging the analytical capabilities of Microsoft Excel to facilitate a structured and organized approach to data collection and analysis. The same tool was used for data organization and representation. Within this app, a comprehensive table was constructed in which the rows were designated to the considered references and data corresponding to individual RQs were entered into separate columns, fostering a systematic representation of the data obtained. Subsequently, to enhance the reliability and validity of the data integration process, a second reviewer performed a verification of the initially extracted data. This encompassing procedure ensured a high degree of accuracy and reduced potential discrepancies, thus guaranteeing the integrity of the data.

Results

As a result of applying the PRISMA guidelines, a total of 28 publications were considered in this study to address the RQs.

RQ 1: What Is the Papers’ Distribution Over Time?

Considering the time range for paper analysis, Figure 4 depicts the time distribution of publications over the period of 2012 to 2022.
Analyzing Figure 4, we can see an increasing trend in the number of publications starting in 2015, with the maximum value being reached in 2019. This denotes an increasingly high interest in these technologies. The number of publications decreased again in 2020, probably because of the pandemic, which limited human interaction and prevented the development and testing of apps dedicated to children with autism.

Regarding the types of publications, more than half (17/28, 61%) were presented at conferences, and 39% (11/28) were articles published in specialized scientific journals.

**RQ 2: What Category of People Are the Studies Aimed at?**

The analysis of the considered publications revealed that 43% (12/28) [29,31-33,36,39,43,44,47,50,51,53] stated that they were about children without mentioning the number of participants or their ages. Textbox 1 summarizes the data obtained.

The paper by Xia et al [47] addressed people with autism without mentioning whether they were children or adults, and the study by Wang et al [54] only involved adults but was of interest because the proposed solution can be applied to children as well.

In addition, the studies by Zheng et al [38], Escobedo et al [46], and Voss et al [49] included children both with and without ASD to compare the results and the process recorded in both cases.
Textbox 1. Number and age of the children involved in the studies.

- Hashim et al [28]: 6 children aged between 5 and 12 y
- Machado et al [29]: children—age not stated
- Tang et al [30]: children aged <4 y and between 4 and 8 y; number not mentioned
- Selvarani et al [31]: children—age not stated
- Abou El-Seoud et al [32]: children—age not stated
- Vullamparthi et al [33]: children—age not stated
- Singh et al [34]: children aged between 9 and 12 y
- Chen et al [35]: 6 teenagers aged between 11 and 13 y
- Tang et al [36]: children—age not mentioned
- Giraud et al [37]: 12 children aged between 5 and 9 y
- Zheng et al [38]: 12 children, 6 with autism spectrum disorder (ASD) and 6 with typical development
- Pradibta and Wijaya [39]: children—age not stated
- Nubia et al [40]: 6 children (5 boys and 1 girl) aged between 3 and 9 y
- Sait et al [41]: 9 children aged between 4 and 12 y
- Wan et al [42]: 10 children aged between 3 and 8 y
- Kavitha et al [43]: children—age not stated
- Silva et al [44]: children—age not stated
- Kalantarian et al [45]: 8 children aged between 6 and 12 y
- Escobedo et al [46]: unknown number of children aged between 8 and 11 y, including 3 children with autism
- Xia et al [47]: mainly people with autism
- Amado et al [48]: children aged between 7 and 9 y; number not indicated
- Voss et al [49]: 20 children with ASD and 20 children without ASD
- Washington et al [50]: 14 families
- Gulati and Handa [51]: children—age not stated
- Escobedo et al [52]: 12 children and 7 teachers
- Bouaziz et al [53]: children—age not stated
- Wang et al [54]: 4 adults, but the system was suitable for children as well
- Gelsomini et al [55]: 5 children (2 with mild ASD, 2 with medium ASD, and 1 with psychomotor retardation)

RQ 3: What Technologies Are Used With XR or Any of Its Subdivisions to Develop Apps for ASD Therapy?

AR mobile apps for therapy for children with ASD typically used a combination of the following technologies:

1. Mobile devices such as smartphones and tablets equipped with cameras; displays; and sensors such as accelerometers, gyroscopes, and GPS
2. AR software development kits such as ARKit, ARCore, and Vuforia, which provide the tools and framework needed for developing AR apps
3. Graphical and game engines such as Unity and Unreal Engine for 3D model development and creating animations and interactive environments
4. Natural language processing and speech recognition technologies for creating voice-activated AR experiences
5. Computer vision and image-processing techniques for real-time object tracking and recognition of objects, faces, and gestures
6. Machine learning algorithms for customizing the AR experience based on the child’s performance and preferences
7. Cloud computing infrastructure for data storage, management, and analysis of therapy progress

Of the 28 analyzed publications, 22 (79%) addressed a solution from the spectrum of AR implemented on mobile devices such as smartphones owing to their processing power and integrated sensors that make them suitable tools for implementing apps without the need for additional and sophisticated equipment. In addition, the papers by Giraud et al [37], Sait et al [41], Gulati and Handa [51], and Gelsomini et al [55] presented solutions based on VR. Although the articles by Wan et al [42] and Kalantarian et al [45] did not present an AR or VR solution, the methodology addressed and the results obtained show the
potential for research in this area. Textbox 2 summarizes information about the technologies and platforms used for app development.

Unity 3D and Vuforia were among the most common platforms used in the development of AR apps for mobile devices, with the Android operating system often mentioned. Several studies (6/28, 21%) [29,38,41,49,51,55] used wearable devices such as Google Glass, Oculus Go, Google Cardboard, Leap motion sensors, and E4 wearable sensors along with the mentioned technologies. Interactive cards were also used as markers to overlay virtual content.

Textbox 2. Technologies and platforms used.

- Hashim et al [28]: interactive cards, augmented reality, and smartphones
- Machado et al [29]: augmented reality based on smart glasses and Android, web platform, Node.js, eye tracker, sensors, and Amazon Alexa
- Tang et al [30]: augmented reality and Google TensorFlow
- Selvarani et al [31]: interactive cards, augmented reality based on markers, Vuforia, Android smartphone, and Unity 3D
- Abou El-Seoud et al [32]: augmented reality based on markers, smartphones, and the Aurasma framework
- Vullamparthi et al [33]: smartphone, Android, augmented reality, and QR codes
- Singh et al [34]: desktop app and augmented reality
- Chen et al [35]: Vuforia and smartphone or tablet PC
- Tang et al [36]: Google TensorFlow, augmented reality, and smartphone or PC
- Giraud et al [37]: virtual reality (VR) and Unity 3D
- Zheng et al [38]: augmented reality, Microsoft Kinect, and portable E4 sensor
- Pradibta and Wijaya [39]: interactive cards, augmented reality, Android smartphone, and Adobe for animation and graphic illustration
- Nubia et al [40]: augmented reality, Android tablet PC, Unity 3D, Vuforia, and Blender
- Sait et al [41]: VR, Unity 3D, and VR glasses (Oculus Go)
- Wan et al [42]: system that can be implemented on a PC, smartphones or robots; no use of augmented reality or VR
- Kavitha et al [43]: augmented reality, Android smartphone, Vuforia, and ARCore
- Silva et al [44]: augmented reality, smartphone or tablet PC, and Vuforia
- Kalantarian et al [45]: Android smartphone; no VR or augmented reality
- Escobedo et al [46]: augmented reality and Android smartphone
- Xia et al [47]: augmented reality, Android or iOS smartphone, React, Node.js, and Python for object recognition
- Amado et al [48]: augmented reality, Vuforia, Unity 3D, Android smartphone, Balsamiq Mockups 3, and Tinkercad
- Voss et al [49]: augmented reality, Android smartphone, and Google Glass
- Washington et al [50]: Google Glass and Android smartphone
- Gulati and Handa [51]: VR, Leap motion sensors, and VR camera
- Escobedo et al [52]: augmented reality, smartphone or tablet PC, PC server, MySQL database, and HTTP
- Bouaziz et al [53]: interactive cards, augmented reality, smartphone, and Vuforia
- Wang et al [54]: augmented reality, tablet PC or smartphone, Unity 3D, and Vuforia
- Gelsomini et al [55]: VR, Google Cardboard, smartphone, and Unity 3D

Table 3 presents the number of publications aimed at improving basic skills.

A total of 25% (7/28) of the publications [28,30,33,36,40,49,52] focused on improving communication skills such as English vocabulary learning [28]; word learning using automatic object recognition through an app based on the TensorFlow library that can be used either when connected to the internet or offline [29]; speaking, reading, and associating images using an app that allows for customization of lessons by parents or therapists [33]; and communication and socialization by delivering certain cues through smart glasses [49]. Textbox 3 details the skills targeted in the studies.

RQ 4: What Skills Were Targeted for Improvement in Primary Studies?

Owing to the deficiencies of children with ASD, the aim was to improve some basic skills such as the following:

1. Communication and language development
2. Social interaction and play skills
3. Fine and gross motor skills
4. Emotional regulation and collaborative strategies
5. Cognitive and problem-solving abilities
6. Attention and ability to follow instructions
7. Independence and self-help capabilities
Table 3. Targeted learning skills (n=28).

<table>
<thead>
<tr>
<th>Skill</th>
<th>Studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Religious skills</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Daily activities, meal preparation, toothbrushing, and eating</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Cognitive or attention</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Expressing emotions or social skills</td>
<td>5 (18)</td>
</tr>
<tr>
<td>Environment adaptation</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Motor skills</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Task training</td>
<td>1 (3)</td>
</tr>
<tr>
<td>General skills</td>
<td>3 (11)</td>
</tr>
<tr>
<td>Number learning</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Object recognition</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Communication or vocabulary</td>
<td>7 (25)</td>
</tr>
</tbody>
</table>

Textbox 3. Skills aimed to be improved.

- Hashim et al [28]: communication skills; learning English vocabulary, pronunciation, and articulation skills
- Machado et al [29]: daily routine activities (preparing meals)
- Tang et al [30]: word learning and object recognition
- Selvarani et al [31]: number learning
- Abou El-Seoud et al [32]: general skills; the user can choose the augmented reality (AR) content to be displayed
- Vullamparthi et al [33]: speaking abilities, reading, image associations, and activity scheduling
- Singh et al [34]: procedural task fulfillment
- Chen et al [35]: expressing emotions and social abilities
- Tang et al [36]: object recognition and vocabulary learning skills
- Giraud et al [37]: motor and social skills
- Zheng et al [38]: toothbrushing abilities
- Pradibta and Wijaya [39]: religious abilities—prayers
- Nubia et al [40]: communication abilities
- Sait et al [41]: adaptation to a new or unfamiliar environment
- Wan et al [42]: cognitive skills and practicing facial emotions
- Kavitha et al [43]: general skills; the user can choose the AR content to be displayed
- Silva et al [44]: social and general skills
- Kalantarian et al [45]: expressing emotions and social abilities
- Escobedo et al [46]: social skills in real-life situations, building and maintaining social relationships, improving conversational ability, and managing behavior and emotions
- Xia et al [47]: social and self-help abilities (shopping)
- Amado et al [48]: cognitive skills
- Voss et al [49]: social and communication abilities
- Washington et al [50]: expressing emotions
- Gulati and Handa [51]: motor, focusing, and general skills
- Escobedo et al [52]: object recognition
- Bouaziz et al [53]: self-help skills (feeding)
- Wang et al [54]: attention skills
- Gelsomini et al [55]: general skills (attention, concentration, and understanding) and narration
Some studies (3/28, 11%) focused on the development of self-help skills such as preparing meals with the help of smart glasses, receiving real-time information about the steps to follow [29], brushing teeth [38], or eating [53]. By improving the skills aimed at in the studies considered in this review and developing skills that can improve the deficiencies present in children with ASD, the social inclusion of children with ASD was pursued.

**RQ 5: What Are the Purposes for Which the Proposed Solutions Were Used?**

Using at least one of the technologies targeted in this review, the solutions presented in these publications were used to assist children with ASD. Developed for use by both therapists and parents at home or in specialized medical centers, these solutions aimed to improve certain fundamental aspects of the lives of children with autism. **Textbox 4** presents information related to the reasons for which the apps were developed.

Hashim et al [28] created an app for the development of children’s English vocabulary, which could potentially be used with other languages as well. The solution proposed in the paper by Machado et al [29] used multiple technologies to allow the therapist to model activities using a web platform and provide hints to users via smart glasses. It also works as an attention-monitoring tool via an eye tracker so that activities can be evaluated and improved. The study by Tang et al [30] addressed the problem of communication through automatic object recognition using a smartphone and display of virtual content (object names) on the screen. This app works either when connected to the internet or offline. A similar approach was observed in the study by Selvarani et al [31], in which children could learn numbers by scanning notebooks using their mobile devices, after which the relevant content was displayed on the screen.
Textbox 4. The purpose of the developed apps.

- Hashim et al [28]: the Areal-Vocab app was developed to help children with autism improve their English vocabulary.
- Machado et al [29]: the aim was to develop an assistive app using augmented reality (AR) based on smart glasses and a visual attention analysis tool to help people with autism in daily tasks by providing complementary information (eg, to pick up the knife and then cut the strawberries). The therapist can model the activity.
- Tang et al [30]: the researchers intended to develop a mobile app for children with autism that could run either connected to the internet or offline that would improve word learning skills by using object recognition.
- Selvarani et al [31]: the aim was to help children with autism learn numbers by scanning an interactive card using the app so that complementary video and audio content is displayed on the screen.
- Abou El-Seoud et al [32]: the aim was to develop a framework to help parents or educators use AR in a personalized way by choosing what type of AR educational content to display over a printed marker representing a familiar cartoon character.
- Vullamparthi et al [33]: a tool was developed that included an interface for parents or educators to scan a QR code and create various lessons and an interface for children. It used the smartphone camera, an Android apk (Android application package), a web page, a database, and Jakarta server pages.
- Singh et al [34]: the paper was a comparative study that aimed to explore the effectiveness of AR in the execution of tasks among less privileged children (who have had minimal interaction with technology), healthy but younger children, and children with autism.
- Chen et al [35]: the researchers developed a Vuforia-based AR app that can be deployed on Android or iOS devices (smartphones or tablets). This app can scan storybooks (with images captured from videos) and overlay relevant content to assist children with autism in expressing and understanding emotions and developing social skills.
- Tang et al [36]: the aim was to develop a tool for children with autism that can recognize objects and display their names.
- Giraud et al [37]: the study aimed to involve children with autism in common actions (moving furniture) by interacting with a virtual character projected on a tactile magnetized surface.
- Zheng et al [38]: the goal was to develop an AR system (Cheerbrush) that could teach children with autism how to brush their teeth considering how important this is to stay healthy and avoid dental procedures. It uses Kinect to capture the user’s movement, a 3D-printed toothbrush to assess brushing skills, a monitor to view the surroundings, and an avatar. It also uses a wristband to assess children’s stress while using the app.
- Pradibta and Wijaya [39]: the aim was to help children with autism learn daily prayers. The goal was to develop an app that contains animated learning materials in the form of daily prayers from the Islamic religion.
- Nubia et al [40]: the aim was to help children with autism communicate better using an app that can identify human-recognizable objects such as animals, fruits, or other common objects and match them with specific sounds.
- Sait et al [41]: the goal was to develop a virtual reality (VR) framework in which the teacher can enter information about the child and prepare scenes that can be watched by a child wearing VR glasses. The main objective was to familiarize children with autism with places such as school, the schoolyard, and the classroom by previously visualizing the environment.
- Wan et al [42]: the aim was to help children with autism recognize, practice, and express emotions such as happiness, sadness, fear, or anger.
- Kavitha et al [43]: the aim was to help children with autism recognize objects or animals by rendering 3D content over certain images.
- Silva et al [44]: the aim was to help reduce the isolation of children with autism by encouraging them to explore the world with the help of an app based on geolocation and AR.
- Kalantarian et al [45]: the goal was to help children with autism learn to express their emotions. Guess what? is an Android mobile app similar to Heads up!, a game in which a parent holds the smartphone with the screen facing the child, the child imitates what they see, and the parent tries to guess the simulated emotion.
- Escobedo et al [46]: the paper describes the design and development of the MOSOCO app, which is a mobile app that provides real-time support and guidance to children with autism in practicing social skills. The app uses AR technology to overlay social hints directly into the child’s real environment, allowing them to practice social skills in real-life situations.
- Xia et al [47]: the app provided step-by-step guidance for people with autism to go shopping by augmenting real shopping scenes using object recognition, barcode reading, and automatic classification.
- Amado et al [48]: the main objective was to develop an AR mobile app to be used by parents of children with autism for their therapy during the pandemic, when human interaction was limited.
- Voss et al [49]: the system described in the paper aimed to help people with autism spectrum disorder improve their social skills by providing discrete real-time social cues via wearable technology. Social cues are provided directly in the wearer’s field of vision using AR technology and are intended to help the wearer navigate social situations and improve their social interactions and communication skills.
- Washington et al [50]: the goal was to develop an app that runs on an Android smartphone (used by a parent) that is connected to a Google Glass device worn by the child. Social cues are delivered to the glasses based on emotions recognized by the mobile app, which also records the session (video only for privacy reasons). The activities are gamelike—catch the smile, guess the emotions, and unstructured activities.
- Gulati and Handa [51]: the aim was to develop a VR game to improve reading, basic math, and spelling. Motor skills are improved by reading gestures and helping coordinate them with the eyes using the Leap motion sensor.
Escobedo et al [52]: the aim was to create an app that can identify objects that are tagged and display relevant content over them, such as text, 3D models, vibrations, video, or audio, and the user can receive a reward. The main architecture is composed of a module called therapy manager, an ambient notification system, and a tag manager.

Bouaziz et al [53]: the aim was to develop an app dedicated to children with autism that teaches them how to eat by scanning an interactive card and displaying on top of it a 3D character depicting the targeted skill.

Wang et al [54]: the aim was to help adults with autism be more focused by performing certain tasks, such as rearranging objects in a scene.

Gelsomini et al [55]: the aim was to develop a VR mobile app for smartphones that can be used with Google Cardboard, helping children with autism understand activities through storytelling and allowing caregivers to customize the content using a web app, monitor children’s attention, and analyze statistics.

The apps developed in the studies by Abou El-Seoud et al [32] and Vullamparthi et al [33] aimed to go through some lessons that parents and educators could customize by accessing a web platform so that they could choose which type of content to display when the app detected an object in the visual area. The study carried out by Singh et al [34] compared the effects of apps that use AR to perform certain tasks in both children with ASD and children with typical development. Religious activities were also included in one study [39], which presented an app containing animated materials that helped children learn prayers.

Given the fact that children with autism typically experience difficulties adapting to a new environment, Sait et al [41] aimed to develop an app that uses VR, VR glasses, and a web platform in which therapists can enter information about each child and set up custom scenes, such as a classroom, to be viewed virtually and get used to. In addition to the goal of conducting basic activities, Xia et al [47] developed an app to guide individuals with autism with grocery shopping step by step.

RQ 6: What Are the Results Obtained Using the Proposed Solutions?

Depending on the proposed solution and the objectives of the studies, the results were different, but in general, where the app was tested, encouraging results were obtained, with the remark that these were to be improved and tested more thoroughly. In cases in which the app was not tested with the intended audience but was proven to function, it was deemed to have potential. Textbox 5 summarizes the results obtained in each study.

Upon analyzing the results of the studies included in this review, it was found that only 68% (19/28) of the apps were tested with children with autism, whereas 32% (9/28) were tested only by the developers for functionality purposes. Regarding the methods used to quantify the results, 54% (15/28) of the studies used interviews, and only 14% (4/28) of the studies used an assessment method based on assigning a score according to the degree of skill improvement after using the apps.
Textbox 5. Summary of the results.

- Hashim et al [28]: children and their parents or educators in the study used the app and reported positive results based on interviews: “Helps listen and understand instructions, helps maintain attention longer, helps with pronunciation and enunciation, helps keep them engaged and interested to learn the vocabulary in depth.”

- Machado et al [29]: the app has great potential considering the fact that smart glasses can very easily transpose the user into the world of augmented reality (AR) and help them by displaying complementary information, as well as giving feedback to the therapist. It has been tested by developers but has not been tested with children with autism, so it does not show quantifiable results.

- Tang et al [30]: the first pilot study was conducted on a university campus with neurotypical children and adults, who provided positive feedback and showed a lot of interest. The second study was conducted in a special education unit involving 2 groups: one with children aged <5 y and one with children aged between 6 and 8 y. It was noticed that the younger children had difficulty using the app, but it was well received by the older children. Positive feedback was also provided by parents and teachers, pointing out that the offline module required improvement.

- Selvarani et al [31]: the app (NUM09) is functional but has not been tested on children with autism with quantifiable results.

- Abou El-Seoud et al [32]: a total of 3 patients with autism, together with their instructors, performed a usability test. According to responses to a questionnaire, the system can improve communication, concentration, and attention and is easy to use.

- Vullamparthi et al [33]: this study developed an Android smartphone app that helps children with autism and their parents or therapists create personalized lessons to improve basic skills such as reading, writing, or picture recognition. A workshop was held, and positive feedback from parents was reported. There are no quantifiable results.

- Singh et al [34]: the main task was to complete a tangram puzzle. In the first stage, the involved children did not solve the puzzle without clues involving AR, but it was reported that solving took longer in the AR training mode. In the first study, children aged 9 to 12 y rated the desktop-based instruction mode as the least preferable, whereas the performance using the AR mode was superior. In the second study, 4 children with autism followed the same procedure but had difficulty using the AR-based solution, resulting in poorer performance on the task.

- Chen et al [35]: the app was tested in a dedicated room equipped with a computer, a 52-inch monitor, and 8 tablets. The therapist showed the children the app and asked them to look at the pictures, answer some questions, and use the tablet to access the AR content by pointing it at the picture with the app running in the background. Positive feedback was reported from the children, who were curious and eager to discover new visual cues, showing interest in the facial expressions, gestures, and related activities of the characters. The children had low scores on the initial assessment, but all 6 scores increased significantly after the app intervention. The most dramatic improvement was in one child, from 30% to 89.5%.

- Tang et al [36]: the app works, but it has not been tested on children with autism with quantifiable results.

- Giraud et al [37]: 12 children with autism spectrum disorder (ASD; including 2 girls) aged between 5 and 9 y participated in the study for a period of 3 mo. In total, 7 of the children showed little conversational language. A 3-stage experiment was conducted (familiarization, moving an object with an agent following the child, and training with an agent that the child follows). Preliminary results were encouraging: one-third of the children completed the training, another third needed device adjustments, and some had difficulty using the system.

- Zheng et al [38]: to evaluate the system, 6 children aged between 3 and 6 y (3 with ASD and 3 without ASD) were involved in an experiment comparing the results. It was noted that all the children were able to complete the training sessions, but the children with ASD were clearly more engaged and interested. After training, the most notable improvements were observed in children with autism. During an interview, both children and parents said that they liked the app and that it helped them improve their toothbrushing skills.

- Pradibta and Wijaya [39]: no proof of testing with children with autism and no quantifiable results.

- Nubia et al [40]: by playing relevant sounds in line with images, the app helped children improve their learning skills compared with traditional methods. A 14% increase in attention and a 9% increase in verbal language were reported.

- Sai et al [41]: the system was used by 9 children with autism who benefited from the help of therapists who guided them in adjusting the Oculus Go headset and using the app (AutiVE). One of the issues was the virtual reality (VR) headset itself and the VR environment, but the website provided had a video explaining them. In total, 8 of the children eventually accepted the device. There were some improvements in learning skills, but no detailed statistics were mentioned.

- Wan et al [42]: the children completed a 20-min training session each day for 4 consecutive days. A total of 6 participants showed improvement in proficiency in operating the system, 5 of 6 completed all tasks, and 4 of 6 showed improvements in expressing emotions. Children aged <5 y found the app difficult and did not perform in a satisfactory manner.

- Kavitha et al [43]: the app works, but it has not been tested with quantifiable results.

- Silva et al [44]: an app similar to Pokémon GO was developed in which users can find “monsters” in certain areas and, by clicking on them, find relevant information. The concept of gamification was used, but the system was not validated with real users with autism.

- Kalantarian et al [45]: the solution was tested with 8 children, all boys, playing up to 5 games in 1 session. In total, 94%, 81%, 92%, and 56% of the emotions were labeled correctly as disgust, neutrality, surprise, and fear, respectively.

- Escobedo et al [46]: the app was evaluated over 7 wk. Interview results revealed that the app was well received by children with autism and their therapists and that it was effective in helping children practice and improve their social skills in real-world situations. The authors reported that users were able to use the app easily and that the AR technology was effective at providing children with real-time support and feedback. The study also showed that the app was well accepted by therapists, who found it a useful tool for their patients’ therapy.

- Xia et al [47]: the app, called ParaShop, was tested by a nonprofit organization that helps people with disabilities. Staff said that the app helped people with autism buy their groceries, but the number of participants or other details were not mentioned.
Amado et al [48]: a case study was conducted using Google Forms asking parents to answer questions related to their children (eg, age, gender, and whether the parents lived together). Several studies with parents were conducted, and then the app was developed based on their responses and requirements. In the last stage of the case study, 5 questions were posed about the final prototype of the app. The survey revealed that 46.2% of parents were satisfied and 23.1% were very satisfied. Overall, the mobile app received positive feedback from respondents.

Voss et al [49]: the research entailed a study involving 20 participants with ASD and 20 participants without ASD who used a system called Superpower Glass over a 4-mo period. The results showed that the participants found the social cues useful in situations and improved their social interactions and communication skills. The study also assessed the acceptability and usability of the system, and the results suggest that it was well received by participants and easy to use.

Washington et al [50]: the app was tested by families, and they reported that it was useful, with some of them recording the sessions and then showing them to the children to see how they behaved for further improvement. Overall, based on interviews, parents reported positive outcomes.

Gulati and Handa [51]: the concept of gamification was used; it has potential, but it has not been tested in children with autism. To play the game, a dedicated gaming room and specific equipment are required.

Escobedo et al [52]: the app (Mobis) was tested with 7 teachers caring for 12 children with autism aged between 3 and 8 y. The researchers conducted weekly interviews with the teachers, keeping in mind that only 3 out of 12 children were able to properly pronounce words. The duration of the observation was 54 h. Participants were reported to find Mobis “exciting, useful, and easy to use.” Students improved their motor skills by focusing the camera on the target. Mobis increased the time that students stayed on task by 20% and motivated them to use the app as they were excited to discover new objects in their environment. Selective attention improved by 62%, and sustained attention improved by 45%. Mobis also induced positive emotions and taught behavioral skills such as tolerance.

Bouaziz et al [53]: no proof of testing with children with autism and no quantifiable results.

Wang et al [54]: the app was developed for demonstrative purposes only; it has not been tested with quantifiable results.

Gelsomini et al [55]: the solution (Wildcard) was tested in a special unit with 5 children with autism during 8 individual therapy sessions. Therapists reported improvements in children’s attention and cognitive skills, but the paper only reported qualitative data. Therapists were excited to be able to customize each VR session and noted that patients embraced the app and found it engaging.

**Discussion**

**Principal Findings**

The analysis revealed an increasing trend in publications starting from 2015, reaching its highest point in 2019 and followed by a decline in 2020, potentially because of the pandemic. Most of the papers (17/28, 61%) were presented at conferences and largely focused on AR solutions (22/28, 79%) for mobile devices to assist children with ASD in enhancing basic skills and fundamental life aspects. Notably, Unity 3D and Vuforia emerged as popular development platforms. Although a substantial percentage of publications (13/28, 47%) did not provide details on participating children, most of the identified participants were aged between 3 and 13 years. Developed for use by both therapists and parents at home or in specialized medical centers, these solutions showed encouraging preliminary results but underscore the necessity for further, more extensive testing, particularly as a significant portion (9/28, 32%) were only developer tested.

**Main Directions of Research**

Upon examining the scientific publications included in our study, several main directions for the use of XR to support children with autism can be identified.

One notable topic is the use of AR in the area of language skills and vocabulary learning in children with autism. Researchers in some studies (3/28, 11%) [28,30,43] focused on the development of AR-based mobile apps that facilitate word learning and object recognition through techniques such as deep learning and automatic object recognition.

Another topic addressed in some studies (3/28, 11%) [29,39,53] was the use of smart glasses or wearable devices to support children with autism in social interactions. These devices provide real-time visual cues and information to enhance communication and social interaction skills.

The use of AR occupational therapy and the development of cognitive skills in children with autism were explored in some studies (3/28, 11%) [31,38,46], which proposed AR-based apps to aid children in learning numbers, teeth-brushing skills, or environmental adaptation skills.

Furthermore, it was stated that the apps specifically designed for children with ASD should be tested with a target group of children, and the results should be quantified in a pertinent manner given that a large part of the findings were obtained through interviews.

Personalization and adaptability are other key aspects of developing mobile apps, as shown in the studies by Wan et al [42], Kalantaririan et al [45], and Washington et al [50]. These publications addressed the development of personalized systems and apps to maximize the therapeutic and educational benefits for children with autism.

Some studies (3/28, 11%) [32,34,52] also examined the use of AR to provide individual support for people with autism and cognitive impairment. These studies proposed AR-based frameworks and approaches to assist individuals with autism in various activities and tasks, such as training in procedural tasks, perception and recognition of facial emotions, or assistance in real-life situations.

In addition, some studies (3/28, 11%) [35,37,55] investigated the use of AR in the context of education and social skill development. These studies focused on the use of interactive books, serious games, or training apps to support children with
autism in understanding and interpreting facial expressions, social cues, and social interactions.

The use of AR in the context of learning in a geographical environment or learning environmental coping skills was addressed in some studies (3/28, 11%) [40,41,47]. These studies proposed AR-based apps to assist children with autism in exploring and learning in a geographical environment or in developing adaptive skills applicable to different situations and contexts.

A relatively small number of studies (6/28, 21%) [34,38,46-48,52] focused on VR-based approaches for apps, indicating a shift toward the adoption of AR owing to its lower cost and greater usability. Using VR, the study by Abou El-Seoud et al [32] evaluated joint action (moving furniture) abilities using a virtual character, and the preliminary results were encouraging, with one-third of the children completing the training despite 7 of them having limited conversational language. Another issue addressed in the study by Tang et al [36] was adaptation to unfamiliar surroundings. Researchers reported improvements in learning abilities but stated that the equipment and VR environment posed the greatest challenges. The concept of gamification was integrated with VR in the study by Escobedo et al [46] to enhance reading, basic mathematics, and spelling. This app required a special room to run. Although this app has great potential, it has not yet been tested in children with ASD. In addition, an app using Google Cardboard and VR was developed to enhance the cognitive and attentional skills of children. Therapists expressed satisfaction with the outcomes as they were able to personalize each session using unique teaching methods.

User Interaction Perspectives

Overview

The interaction of children with ASD with XR devices, such as AR and VR platforms, brings forth a distinct set of considerations. The manner in which children with ASD use these devices can be influenced by their sensory sensitivities, motor skills, cognitive abilities, and preferences. Although the experiences can vary widely, the following are some ways in which children use XR devices and the challenges they may face.

Physical Interaction

Children use XR devices by interacting with touch screens, controllers, or wearable components. They may tap, swipe, or perform gestures to navigate through XR environments. However, children with fine motor difficulties may struggle with precise interactions, leading to accidental inputs or difficulties in selecting desired options.

Visual Engagement

Children engage visually with the XR content displayed on screens or through headsets. Visual stimuli can capture their attention and spark interest. Nonetheless, those with sensory sensitivities may experience sensory overload or visual discomfort if the content is excessively bright, flashy, or overwhelming.

Spatial Awareness

XR experiences often involve spatial interactions such as moving through virtual environments or manipulating virtual objects. Children’s spatial awareness skills can influence their ability to navigate these environments. Some children may find it challenging to grasp the concept of a virtual space, leading to disorientation.

Auditory Response

Many XR apps incorporate auditory cues, sound effects, or voice instructions. Children may respond to auditory prompts by vocalizing or reacting physically. However, children who are sensitive to loud or sudden sounds may experience distress in XR environments with intense auditory stimuli.

Attention and Engagement

Children’s level of attention and engagement with XR content can vary. Some may become deeply immersed and engaged, whereas others may have difficulty sustaining their attention because of the novelty of the experience or sensory distractions.

Preferences and Comfort

Children’s preferences for certain types of interactions or content can influence their engagement. Some children may appreciate exploring virtual worlds, whereas others may prefer more structured or repetitive activities. Ensuring a variety of XR experiences allows for accommodating different preferences.

Transition Challenges

Transitioning between the real world and the XR environment can be challenging for some children. They may have trouble understanding that the virtual elements are not physically present or struggle with transitioning back to reality after prolonged XR use.

Response Variability

Children with ASD may respond to XR experiences differently across sessions. Factors such as mood, sensory sensitivities, and cognitive states can influence their interactions. Some days, children may be more receptive to XR, whereas on other days, they may be less engaged or overwhelmed.

Calibration and Setup

XR devices require proper calibration and setup for optimal interaction. Children may need assistance in adjusting headsets, ensuring proper alignment, or calibrating controllers. Technical difficulties can lead to frustration or disengagement.

Challenges

The identified challenges are as follows:

1. Individualized learning needs: a prevalent challenge across the studies in this review was catering to the diverse learning preferences and abilities of children with ASD. For instance, Hashim et al [28] faced the task of addressing the specific needs of children with mild ASD. Similarly, the studies by Abou El-Seoud et al [32] and Singh et al [34] addressed the challenge of tailoring their AR experiences to suit varying preferences and capabilities.
2. Transferability and generalization: a common limitation is the transfer of learned skills to real-world scenarios. As seen in the studies by Tang et al [36] and Giraud et al [37], researchers have encountered challenges in translating acquired skills into practical applications. In addition, studies such as those by Chen et al [35] and Kavitha et al [43] noted limitations in transferring learned skills beyond the AR context, possibly owing to the variations in real-world stimuli.

3. Technical feasibility and personalized support: technical feasibility and ongoing support emerged as challenges in some studies [29,38]. Maintaining the functionality of AR-based smart glasses and ensuring accurate real-time feedback for toothbrushing techniques required continuous technical support.

4. Sensory overload and individualization: sensory sensitivities and the need for individualized solutions were prominent challenges. Sait et al [41] encountered the challenge of designing virtual environments that cater to sensory sensitivities, whereas studies such as the one by Voss et al [49] highlighted the importance of unobtrusive cue presentation in wearables for children with ASD.

5. Cognitive adaptation and user adoption: cognitive adaptation and user adoption challenges were evident in some studies [48,52]. Designing tasks that effectively target cognitive skills and maintaining user engagement over time were key considerations.

6. Ethical implications of data handling: as AR interventions involve interactions and data collection, ethical considerations are paramount. The study by Wan et al [42] delved into recognizing facial expressions, which raises ethical concerns related to data privacy and security. Ensuring that data-handling protocols adhere to ethical standards becomes crucial, underscoring the need to protect sensitive user information while deriving meaningful insights from the interactions.

7. Cross-cultural adaptation and applicability: given the diversity of cultures and languages, ensuring the cross-cultural adaptation and applicability of AR interventions becomes a notable challenge. The study by Wang et al [54], which explored mobile AR for attention improvement in adults with ASD, highlights the importance of adapting interventions to diverse cultural contexts. This challenge emphasizes the need for cultural sensitivity and the localization of content to ensure that interventions are universally accessible and effective.

8. Long-term impact measurement: measuring the long-term impact of AR interventions and tracking the progress of children over time poses significant challenges, as pointed out in the study by Escobedo et al [46], which emphasized the importance of assessing the sustained effects of interventions beyond short-term interactions. This challenge underscores the necessity of devising reliable methodologies for gauging the lasting benefits of AR interventions and understanding how these interventions contribute to the developmental trajectory of children with ASD.

In the realm of AR apps for children with ASD, studies have striven to engage children through various interaction modes while tackling shared challenges. The diverse engagement strategies and the collective endeavor to overcome common limitations underscore the continuous efforts to create meaningful and effective AR-based interventions for this unique demographic.

Research in this area demonstrates an interdisciplinary approach involving collaboration among specialists in education, IT, and mental health. This is illustrated by the diversity of authors and publications included in this review, suggesting that integrating AR, VR, and MR into ASD pedagogy requires a comprehensive approach that considers multiple aspects—from technology design to educational and mental health psychology. It was also specified that the apps aimed at children with ASD should be tested with a target group of children and that the results should be quantified in a more relevant manner given that a large part of the reported results was obtained only through interviews. The analysis of the studies indicates a trend in research toward the use of diverse and innovative study methods, such as using both quantitative and qualitative methods to investigate the impact of AR, VR, and MR on people with ASD.

Furthermore, the analyzed publications suggest that the development and implementation of AR-, VR-, and MR-based technologies extend beyond academia or research, involving partnerships with the private sector and local communities. This demonstrates the awareness of the need to transfer research findings into practice to have a direct impact on people with autism.

**Limitations**

Considering the publications reviewed in this study, several limitations were identified regarding the development and testing of AR-, VR-, and MR-based mobile apps:

1. **Sample size**: some studies involved small samples of participants, which may have limited the generalizability of their results to a larger population of children with autism. The involvement of a limited number of participants in many studies can be attributed to the unique characteristics of the target population—children with ASD. The diversity in ASD manifestation, severity, and individualized needs necessitates careful participant selection. Moreover, recruitment challenges, ethical considerations, and the resource-intensive nature of working with children with ASD contribute to the small sample sizes. However, this limitation was often acknowledged in the papers, along with the understanding that the findings may not be easily generalizable to the broader population with ASD.

2. **Study duration**: the duration of the studies included in this review varied from short testing sessions to several weeks or months. The short duration of many studies was due to practical constraints and the inherent complexities of conducting research involving children with ASD. Longitudinal studies involving children with ASD can present challenges in terms of participant retention, compliance, and data collection consistency over extended periods. In addition, the rapid pace of technological advancements may affect the relevance of the findings if studies are conducted over prolonged durations. However, the researchers did recognize the limitations imposed by...
short study durations and provided justifications for the chosen time frames.

3. Diversity of diagnosis and level of functioning: autism is a disorder with a wide variety of symptoms and levels of functioning, which adds complexity and variability to the research.

4. Standardized outcome assessment: some studies did not use standardized outcome assessment tools, which may have affected the comparability and validity of the obtained results.

5. Availability and accessibility of technology: although the presented studies demonstrate the potential of AR technology to support children with autism, it is important to consider the availability and accessibility of this technology in real-world settings. The cost, infrastructure, and availability of AR devices and apps may be limiting factors in the widespread adoption of this technology.

The relatively small sample sizes and short durations commonly observed in many studies involving children with ASD and XR interventions are notable aspects of the research landscape. Although solutions to these issues were not always addressed in the papers, they remain ongoing areas of consideration for researchers in the field. In the selected papers, although some discussions and considerations regarding the challenges of small sample sizes and short study durations were present, comprehensive solutions were not elaborated on. The researchers often acknowledged these limitations and offered potential insights or recommendations, but definitive solutions were not always a primary focus of the papers, the primary focus being on the technical details and potential outcomes of their approach.

In addition, this study itself has several limitations, which should be considered for further research:

1. Limited number of databases queried: despite using comprehensive search strategies, it is possible that some relevant studies were omitted because they were published in nonindexed or less accessible sources.

2. Field evolution: this field of study is rapidly evolving, and new research may have been published since the literature search was conducted. Consequently, this review may not capture the most recent evidence and emerging trends in the field.

3. No distribution of publication authors: this review did not present information on the regional distribution of authors or the origin of the apps and systems, disregarding the influence of cultural differences on the development of these types of apps.

4. Lack of security analysis: this study did not analyze the security issues associated with the proposed solutions.

5. Absence of cost information: no information regarding the cost of the presented solutions could be identified.

Conclusions

This study aimed to conduct a systematic review of the specialized scientific literature in terms of applications, devices, and technologies relevant to the development of AR-, VR-, and MR-based mobile apps dedicated to the therapy of children with ASDs, an objective that was successfully achieved. At the beginning of this paper, the general concept of ASD was presented, after which the RQs and inclusion and exclusion criteria were defined and the results of applying the PRISMA guidelines for the selection of publications to be reviewed were reported. The answers to the RQs were discussed. At the end of the paper, the limitations of the research were presented. Although the concepts of AR, VR, and MR are not entirely new, their use in the development of therapeutic apps for children with autism has only recently gained popularity. The findings documented in various publications indexed in 5 scientific databases emphasize the suitability of these technologies for such therapy, thereby warranting further in-depth research and the future development of apps based on these technologies. The studies indicated a clear trend toward the use of AR, VR, and MR technologies as a pedagogical tool for people with ASD. This trend involves multidisciplinary collaborations and an integrated approach to research, with a focus on empirical evaluations and ethics regarding the use of technologies. As the field advances, it is essential that research and practice continue to be guided by a balanced and integrated approach that considers both the technological possibilities and the needs and rights of individuals with ASD. However, there are still many issues that require further exploration and research.

Moreover, the publications studied illustrate a wide range of research areas related to the use of AR, VR, and MR in the context of ASD, as well as a variety of methodological and theoretical approaches adopted by the researchers. This suggests that the field is in a phase of rapid growth and diversification, with a wealth of opportunities for future research and development.

Acknowledgments

This research was funded by the 119722/Centru pentru transferul de cunoștințe între întreprinderi din domeniul ICT—CENTRIC, Contract subsidiar 22080/05.10.2022/Autism/ASSIST, Autism ASSISTant—Asistent virtual pentru dezvoltarea abilităților cognitive ale copiilor cu patologie de spectru autist project, contract 5/AXA 1/1.2.3/G/13.06.2018, code SMIS 2014+ 119722 (ID P_40_305).

Data Availability

All data generated or analyzed during this study are included in this published article (and its multimedia appendices).
References


20. Milgram P, Kishino F. A taxonomy of mixed reality visual displays. IEICE Transact Inf Syst 1994 Dec;E77-D(12):1321-1329 [FREE Full text]


https://games.jmir.org/2024/1/e49906


Abbreviations
- AR: augmented reality
- ASD: autism spectrum disorder
- MR: mixed reality
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
- QC: quality criterion
- RQ: research question
- VR: virtual reality
- XR: extended reality

Please cite as:
Toma MV, Turcu CE, Turcu CO, Vlad S, Tiliute DE, Pascu P
Extended Reality–Based Mobile App Solutions for the Therapy of Children With Autism Spectrum Disorders: Systematic Literature Review
JMIR Serious Games 2024;12:e49906
URL: https://games.jmir.org/2024/1/e49906
doi:10.2196/49906
PMID:38373032
Immersion Virtual Reality–Based Methods for Assessing Executive Functioning: Systematic Review

Rebecca Kirkham¹, BPsy (Hons); Lars Kooijman², MSc; Lucy Albertella¹, PhD; Dan Myles¹, PhD; Murat Yücel³,4*, PhD; Kristian Rotaru¹,5*, PhD

¹Turner Institute for Brain and Mental Health, School of Psychological Sciences and Monash Biomedical Imaging Facility, Monash University, Clayton, Australia
²Institute for Intelligent Systems Research and Innovation, Deakin University, Geelong, Australia
³Queensland Institute of Medical Research Berghofer Medical Research Institute, Herston, Australia
⁴Department of Psychiatry, School of Clinical Sciences, Monash University, Clayton, Australia
⁵Monash Business School, Monash University, Caulfield, Australia

* these authors contributed equally

Corresponding Author:
Rebecca Kirkham, BPsy (Hons)
Turner Institute for Brain and Mental Health
School of Psychological Sciences and Monash Biomedical Imaging Facility
Monash University
770 Blackburn Road
Clayton, 3168
Australia
Phone: 61 (03) 9905 0100
Email: rebecca.kirkham@monash.edu

Abstract

Background: Neuropsychological assessments traditionally include tests of executive functioning (EF) because of its critical role in daily activities and link to mental disorders. Established traditional EF assessments, although robust, lack ecological validity and are limited to single cognitive processes. These methods, which are suitable for clinical populations, are less informative regarding EF in healthy individuals. With these limitations in mind, immersive virtual reality (VR)–based assessments of EF have garnered interest because of their potential to increase test sensitivity, ecological validity, and neuropsychological assessment accessibility.

Objective: This systematic review aims to explore the literature on immersive VR assessments of EF focusing on (1) EF components being assessed, (2) how these assessments are validated, and (3) strategies for monitoring potential adverse (cybersickness) and beneficial (immersion) effects.

Methods: EBSCOhost, Scopus, and Web of Science were searched in July 2022 using keywords that reflected the main themes of VR, neuropsychological tests, and EF. Articles had to be peer-reviewed manuscripts written in English and published after 2013 that detailed empirical, clinical, or proof-of-concept studies in which a virtual environment using a head-mounted display was used to assess EF in an adult population. A tabular synthesis method was used in which validation details from each study, including comparative assessments and scores, were systematically organized in a table. The results were summed and qualitatively analyzed to provide a comprehensive overview of the findings.

Results: The search retrieved 555 unique articles, of which 19 (3.4%) met the inclusion criteria. The reviewed studies encompassed EF and associated higher-order cognitive functions such as inhibitory control, cognitive flexibility, working memory, planning, and attention. VR assessments commonly underwent validation against gold-standard traditional tasks. However, discrepancies were observed, with some studies lacking reported a priori planned correlations, omitting detailed descriptions of the EF constructs evaluated using the VR paradigms, and frequently reporting incomplete results. Notably, only 4 of the 19 (21%) studies evaluated cybersickness, and 5 of the 19 (26%) studies included user experience assessments.

Conclusions: Although it acknowledges the potential of VR paradigms for assessing EF, the evidence has limitations. The methodological and psychometric properties of the included studies were inconsistently addressed, raising concerns about their validity and reliability. Infrequent monitoring of adverse effects such as cybersickness and considerable variability in sample

https://games.jmir.org/2024/1/e50282

JMI Serious Games 2024 | vol. 12 | e50282 | p.243
(page number not for citation purpose)
Executive functioning (EF) has long been a focus of neuropsychological assessment because of the significant role it plays in everyday functioning. EF is an umbrella term for higher-order cognitive skills used to control and coordinate a wide range of mental processes and everyday behaviors [1-5], including “…mentally playing with ideas; taking the time to think before acting; meeting novel, unanticipated challenges; resisting temptations; and staying focused” [6]. Although a universally accepted definition of EF does not exist [5], there is agreement on the attributes of 3 core executive functions: inhibition, cognitive flexibility, and working memory [2,4,6]. These core executive functions support other higher-order executive functions such as reasoning, planning, and problem-solving [6-8]. As EF impairment has been linked to a variety of mental disorders [9], it is often considered a transdiagnostic risk factor [10].

Although traditional methods used to assess EF are popular [11,12] and well validated [13], they have been criticized for their lack of ecological validity [14,15]. Ecological validity, within the scope of this study, is defined as the “functional and predictive relationship between the person’s performance on a set of neuropsychological tests and the person’s behavior in a variety of real world settings” [16]. Specifically, we interpret ecological validity as comprising 2 principal components: representativeness—the degree to which a neuropsychological test mirrors the demands of a person’s daily living activities that it aims to evaluate [17], sometimes referred to as verisimilitude [18]—and generalizability—the extent to which test performance predicts an individual’s functioning in their daily living activities [17], also known as veridicality [18].

Traditional assessments tend to take a “construct-led” approach, with each test intended to isolate a single cognitive process in an abstract measure. This process of abstraction may limit the ecological validity of the measure by resulting in poor alignment between the test outcomes and real-world functioning. In turn, this produces a large amount of variance in EF that is unaccounted for by traditional tasks. For example, Chaytor et al [19] noted that traditional EF tests accounted for only 18% to 20% of the variance in the everyday executive ability of participants. This lack of explained variance may be attributed to the nature of the testing environment, the constructs assessed in isolation, the participant’s affective state, and the compensatory strategies available to the participant [19]. A related methodological issue, known as the “task impunity problem” [4,20], indicates that the score on an EF task usually reflects not only the systematic variance attributable to the specific aspect of EF targeted by that task but also the (1) systematic variance across multiple types of EF tasks, (2) systematic variance attributable to non-EF aspects of the task, and (3) nonsystematic (error) variance (see the study by Snyder et al [10] for a detailed review). Outside the testing environment, the process of making a decision or planning and eliciting goal-directed behavior in everyday life is often highly dynamic and influenced by numerous internal and external factors [13,14]. Therefore, an ecologically valid assessment tool will need to include relevant contextual, dynamic, and multidimensional features such as affect and physiological state, which traditional assessments cannot include.

Furthermore, although traditional EF assessment tools may be appropriate for clinical populations, they generate less information about functioning in relatively healthy individuals. For example, the Trail-Making Test (TMT) has routinely been administered as a neuropsychological assessment of driving performance. Although some studies have demonstrated a relationship between the two [21,22], others have shown no relationship [23], particularly in nonclinical populations [24,25]. Thus, although traditional tools are adequate for detecting more severe EF impairments, they are less effective in detecting subtle changes in EF and early decline. Increased test sensitivity to detect subtle intra-individual changes may enable better detection of the prodromal stages of cognitive decline. Early detection is important as it enables early intervention, which may in turn improve prognosis. For example, sensitive detection can identify the prodromal stages of Alzheimer disease in seemingly healthy individuals [26] and mild cognitive decline up to 12 years before clinical diagnosis [27]. Similarly, in a situation in which an individual requires a capacity assessment for an activity, traditional assessments may have limited utility for nonclinical populations. The triangulation of multiple data sources such as biosensors may increase sensitivity to better identify subtle changes in capacity.

To address the shortcomings of poor ecological validity and test sensitivity, research on psychological assessment has begun to investigate virtual reality (VR) technology as a means of providing a more naturalistic environment for evaluating EF in clinical neuropsychological assessments. VR enables the development of custom-designed simulated environments that can replicate real-life environments, potentially increasing its ecological validity through representativeness. In addition, VR could increase engagement [28,29], reduce test time, and better integrate data from biosensors with in-task events that facilitate...
assessment. The following sections will expand on these points and consider the importance of validating and assessing the reliability of VR for EF assessment.

**Ecological Validity and Representative Tests**

There is an increasing emphasis on conducting EF assessments using tasks that resemble situations experienced in everyday life [30]. For example, the Multiple Errands Test (MET) [31] requires individuals to run errands in a real environment (eg, a shopping center). Empirical assessment of the MET has demonstrated its generalizability to daily functioning [32] and carer reports of daily functioning [33]. However, given that the MET is designed to be performed in real-life locations, it is impractical for routine administration by clinicians [34,35] and susceptible to the variable features of real-world environments that are outside experimental control. VR can mitigate these difficulties by maintaining the real-world environment without requiring travel while enabling fine-tuned control and uniform presentation of environmental characteristics [36]. Several studies [37-39] have investigated and developed platforms for this purpose, commonly known as the virtual MET.

**Engagement**

VR has the potential to enhance individual engagement more effectively than traditional pencil-and-paper or computerized tasks by offering a fully immersive experience [40]. Recognized as a crucial aspect of cognitive assessment, engagement can be improved through gamification, thereby improving task performance [41]. “Serious games,” defined as games intended for a variety of serious purposes, such as training, learning, stimulation, or cognitive assessment [42], have been shown to be more engaging than nongamified tasks [43-45]. The unique immersive environment of VR captures increased attention, leading to reduced average response times and response time variability [46]. Notably, recent studies using electroencephalography (EEG)-based metrics have shown greater attention elicited in immersive VR paradigms than in 2D computerized assessments [46]. This heightened immersion and engagement in VR may enhance the reliability of the measures by capturing a more accurate representation of an individual’s best effort.

**Cybersickness**

Despite their increased engagement, VR paradigms have the potential to induce cybersickness, which can threaten the validity of the paradigm. Cybersickness (ie, dizziness and vertigo) is akin to motion sickness but occurs in response to exposure to VR [47]. Previous research suggests that there is a negative relationship between cybersickness and cognitive abilities. For example, Nalivaiko et al [47] found that reaction times were moderately correlated (r=0.5; P=.006) with subjective ratings of nausea. Similarly, Sepich et al [48] found that participants’ accuracy on n-back task performance was weakly to moderately negatively correlated (r=-0.32; P=.002) with subjective cybersickness ratings. Therefore, there is reasonable concern that the potential benefits of engagement and ecological validity may be compromised if participants experience cybersickness.

**Validity, Reliability, and Sensitivity**

Arguably, the biggest threat to the utility of VR platforms is that many studies do not document their validity and reliability. A meta-analysis showed that VR assessment tools are moderately sensitive to cognitive impairment across neurodevelopmental, mental health, and neurological disorders [49], demonstrating their promising application in clinical settings. Borgnis et al [50] reviewed the VR-based tools for EF assessment that are currently available, illustrating the plethora of platforms developing in this field. The works by Negu et al [49] and Borgnis et al [50] highlight the utility of VR assessment tools to detect dysfunction and present the various tools in the literature created to investigate EF. Kim et al [51] provided an overview of the research trends using VR for neuropsychological tests and documented the cognitive functions assessed in each study. However, to the best of our knowledge, there is no overview or examination of the psychometric properties of these VR tools or how they are being evaluated.

Typically, novel measures and assessments are validated against current gold-standard tasks for concurrent validity [52]. Concurrent validity can be a reliable means of determining whether two assessments measure the same construct. However, concurrent validity can also occur when two tests contain the same problems, such as inaccurately measuring a particular construct in the same way. Sequentially, many VR tasks are being created from a “function-led” perspective but validated against “construct-led” tasks [53,54]. Given their different approaches, function-led and construct-led assessments should be validated in different ways or at least using several validation approaches. If function-led VR assessments improve upon the validity of current assessment methods, validation techniques may also need to go beyond comparisons with traditional assessments. For example, function-led VR assessments may be better validated against additional alternative methods, such as carer reports, real-life performance (eg, self-care, residence, transportation, and employment), and diagnostic trajectory [49] as opposed to validation through traditional (construct-led) assessment. Without incorporating tests of ecological validity, the potential advantages of VR may go unrecognized. Given the increasingly rapid development of VR neuropsychological assessments, it will be imperative to maintain high validation standards for these tools [55].

Establishing the reliability of novel VR EF assessments is also critical to the integrity of the outcomes. Reliability ensures that the measure yields consistent and repeatable results, a foundational element for test validity. Consequently, both reliability and validity ought to be evaluated for each measurement tool. Test-retest reliability, confirming consistency over time, should be accompanied by the interval between assessments and the correlation of the results. Internal consistency, typically measured using the Cronbach α, should also be reported for each target construct or domain of assessment. Importantly, for immersive VR EF assessments that evaluate multiple EF constructs, it is essential to report the α for each distinct construct rather than a collective coefficient. This is because the coefficient is intended to evaluate item consistency within a scale measuring a single construct; applying
it across disparate constructs could be confusing and potentially misleading.

Consistency of Terminology

Finally, to ensure psychometric precision and build on previous research, EF assessment paradigms must adopt consistent terminology for their target assessment constructs. This field of EF, although of significant interest to both researchers and clinicians, is marked by varied terminology for identical constructs. This issue, longstanding in EF research (see the study by Suchy [5]; for a review, see the study by Baggetta and Alexander [56]), presents challenges to VR in the EF assessment field. For instance, inconsistent terminology hinders the synthesis of research findings. Diverse labels such as “impulsivity” and “impulse control” might, upon examination, refer to the same underlying construct. Consequently, researchers aiming to extend the literature on “impulsivity” might overlook pertinent studies or exclude valuable references because of terminological discrepancies.

This literature review sought to examine and discuss the development of the VR tools used to assess EF with a specific focus on evaluating their psychometric properties. The studies selected for inclusion in this review were those that developed assessment tools for EF either holistically or in part. The aims of this review were to (1) determine the components of EF assessed using VR paradigms, (2) investigate the methods used to validate VR assessments, and (3) explore the frequency and efficacy of reporting participants’ immersion in and engagement with VR for EF assessment.

Methods

Our review methodology followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [57]. In line with the literature, EF was defined as a set of executive functions, including inhibition, cognitive flexibility, and working memory [2,4,6], that support other higher-order executive functions, such as reasoning, planning, and problem-solving [6,8].

Inclusion Criteria

Before conducting the literature search, the inclusion criteria were established. First, only peer-reviewed articles and conference proceedings (complete manuscripts) written in English would be included. Second, articles that detailed an empirical, clinical, or proof-of-concept study in which an immersive virtual environment (ie, using a head-mounted display, not a 2D computer screen) was reported to broadly investigate EF or higher-order cognition or that examined EF via a selection of one or more subconstructs (eg, inhibitory control and working memory) would be included. Finally, only articles with an adult participant population published after 2013 would be included. This temporal limit was based on the release date of the Oculus Rift Development Kit 1 as it was one of the first accessible products for public use of VR. Articles were identified through the EBSCOhost, Scopus, and Web of Science (WoS) citation databases. Scopus and WoS were chosen because of their prominence as citation databases [58]. To compensate for the bias toward engineering and natural science articles found through Scopus and WoS [59], EBSCOhost was searched for articles published in fields such as (clinical) psychology and medicine.

Search Strategy

Keywords were developed by identifying 3 main components that the relevant literature should include. The 3 components were based on “Virtual Reality,” “Neuropsychological Tests,” and “Executive Function.” It was decided not to search for specific components of EF because of the lack of consensus in the field regarding its components. Rather, it was assumed that, if an article addressed EF or a component of EF, it would include “executive functioning” as a keyword in the title, abstract, or keywords. Other reviews looking broadly at VR paradigms have used similar search strategies [49].

In this study, key terms were developed by identifying synonyms for key components and concatenating them using the “AND” Boolean operator. The final keywords used for the search were as follows: ("virtual" OR “artificial” OR “simulated”) AND ("realit*" OR “world” OR “environment”) AND (neuropsych* OR function* OR cognit*) AND ((executive AND function*) OR (high* AND order AND cognit*)) AND [assessment].

Literature queries made through EBSCOhost were limited to the following databases: Academic Search Complete, AgeLine, AMED, Applied Science and Technology Source, CINAHL, E-Journals, Health Source Consumer and Nursing/Academic Edition, MEDLINE, Mental Measurements Yearbook, Psychology and Behavioral Sciences Collection, and all variations of the American Psychological Association databases. Furthermore, for the search, 3 topic fields (ie, title, abstract, and subject terms) were used to paste the keywords. The 3 topic fields were concatenated using the “OR” Boolean operator. Using the Scopus database, we implemented a basic search in the article title, abstract, or keywords using the keywords. No additional limitations were applied. Our search in WoS included all databases, and the advanced search method was used wherein keyword searches in the article title, abstract, and keyword topic fields were concatenated using the “OR” Boolean operator (ie, Title=(keywords) OR Abstract=(keywords) OR Keywords=(keywords)).

The results for each database were exported to Covidence systematic review software (Veritas Health Information) [60], which removed duplicates. All abstracts were screened independently by the first author and the senior author to determine whether the contents met the inclusion criteria. Full-text screening was also performed by the same authors. Any disagreement was discussed by the first (RK), second (LK), and senior (KR) authors.

Data Extraction

The first and second authors completed the data extraction process by manually reviewing each manuscript; data items (see the following section) were recorded in a tabular format using Microsoft Excel (Microsoft Corp).

Data Items and Synthesis

Demographic details, qualitative descriptions of the VR paradigm, user experience, cybersickness, immersion and...
engagement details, and comparative measures for validation purposes were extracted (Multimedia Appendix 1 [53-55,61-76]).

A qualitative evaluation of the studies included in the review was performed, meaning that the content of each manuscript was assessed based on the reported target constructs or constructs relevant to EF and the extent to which the reported VR task was related to the assessment of the target construct or constructs. To do this, studies were categorized based on the construct they targeted through their VR paradigm as reported by the authors of the respective articles. If multiple constructs were assessed in a single study, the study was included for each construct. No inferences were made about which cognitive construct or constructs was assessed based on the tasks that were reported in the manuscripts. For example, if an article indicated only that they used a VR version of the Stroop test (ST) but did not disclose which construct it assessed using this test, the study was not categorized under inhibitory control or cognitive flexibility but under the general factor “executive functioning.”

Next, it was indicated whether the articles explicitly or implicitly disclosed the way in which the comparative measures (such as particular metrics) were used to validate the VR paradigm. For instance, if the article directly stated a priori that they hypothesized a correlation between a VR task measuring inhibition and a validation task such as the ST, this was recognized as providing explicit validation for inhibition. Conversely, if an article indicated that participants completed the ST, which assessed inhibition and processing speed, and mentioned that the VR paradigm evaluated inhibition, it was considered to provide implicit validation for inhibition. Furthermore, traditional construct- and function-led assessments were identified from the text.

The (quantitative) results of the studies were screened to identify (1) the direction and strength of the relationship between traditional and VR assessments and (2) whether the results from all possible and a priori–defined comparisons were reported.

Finally, qualitative and quantitative tools used to evaluate beneficial and adverse effects of VR immersion were identified from the manuscripts and categorized in a tabulated format. The results of the studies were screened to identify whether they assessed the influence of the beneficial and adverse effects of VR immersion on task performance.

### Results

#### Overview

Through WoS, EBSCOhost, and Scopus, 892 items were identified, from which the Covidence systematic review management platform [60] filtered 337 (37.8%) duplicates. A total of 555 unique articles remained, of which 424 (76.4%) were deemed irrelevant through abstract screening. The final 131 articles had their full texts screened, and 19 (14.5%) met the inclusion criteria. The systematic literature search process is shown in Figure 1.

**Figure 1.** Systematic review process and results from literature searches in EBSCOhost, Scopus, and Web of Science databases.
General EF

In total, 7 of the 19 (37%) of the reviewed studies assessed EF in general, meaning that the authors of these articles did not explicitly state which subconstruct of EF was targeted using the VR task. Table 1 shows which validation tasks were used in each study to measure EF.

Table 1. The validation tasks, authors, and total number of studies examining general executive functioning.

<table>
<thead>
<tr>
<th>Executive functioning: general</th>
<th>VR target construct and validation task</th>
<th>Validation</th>
<th>Authors</th>
<th>Studies examining the construct, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-KEFS&lt;sup&gt;b&lt;/sup&gt; [77]</td>
<td>Implicit</td>
<td>Banville et al [61]&lt;sup&gt;i&lt;/sup&gt;</td>
<td>7 (37)</td>
</tr>
<tr>
<td></td>
<td>TMT-A&lt;sup&gt;c&lt;/sup&gt; and TMT-B&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ST&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modified version of the SET&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HTT&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ZMT&lt;sup&gt;h&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>Implicit</td>
<td>Davison et al [62]&lt;sup&gt;j&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMT-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMT-B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMT-B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OTS&lt;sup&gt;k&lt;/sup&gt;CANTAB&lt;sup&gt;l&lt;/sup&gt;</td>
<td>Explicit</td>
<td>Miskowiak et al [63]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VFT&lt;sup&gt;m&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMT-A</td>
<td>Explicit</td>
<td>Pallavicini et al [64]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMT-B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Groton Maze Learning Test (Cogstate)</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>None specifically reported</td>
<td>N/A&lt;sup&gt;n&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>None specifically reported</td>
<td>N/A</td>
<td>Tsai et al [67]</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>VR: virtual reality.
<sup>b</sup>D-KEFS: Delis-Kaplan Executive Function System.
<sup>c</sup>TMT-A: Trail-Making Test version A.
<sup>d</sup>TMT-B: Trail-Making Test version B.
<sup>e</sup>ST: Stroop test.
<sup>f</sup>SET: Six Elements Test.
<sup>g</sup>HTT: Tower of Hanoi test.
<sup>h</sup>ZMT: Zoo Map Test.
<sup>i</sup>The VR task was predominantly a sorting task for executive functioning assessment. The comparative assessments that validated this assessment were detailed under “executive function” broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.
<sup>j</sup>The VR task was reported to assess executive functioning. The comparative assessments that validated this assessment were detailed under “executive function” broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.
<sup>k</sup>OTS: One Touch Stockings of Cambridge.
<sup>l</sup>CANTAB: Cambridge Neuropsychological Test Automated Battery.
<sup>m</sup>VFT: verbal fluency test.
<sup>n</sup>N/A: not applicable.

Banville et al [61] immersed participants in a Virtual Multitasking Test (VMT), which was in principle designed to measure prospective memory and executive functions by having participants perform multiple tasks in a virtual apartment. However, this paper reported specifically on the task in which participants had to store groceries as fast as possible while also being attentive to other tasks, such as answering the phone or closing a window. Although the authors hypothesized that VMT scores would be correlated with neuropsychological assessments, such as mental flexibility, planning, and inhibition, it was not explicitly stated which metric of the VMT would be correlated with which neuropsychological assessment. Nonetheless, the authors identified that grocery storing time was correlated with the rule-break score on the Six Elements Test \( r_{19} = -0.49; P = .04 \); \( P \) value as reported in the manuscript. Furthermore, the number of errors in storing fruits and vegetables was found to correlate with the perseveration score on the Zoo Map Test \( r_{20} = 0.53; P = .02 \); \( P \) value as reported in the manuscript) and reading speed.
during the second condition of the ST ($r_{20}=0.44; P=0.05; P$ value as reported in the manuscript).

Davison et al [62] immersed participants in a parking simulator and a chemistry laboratory where they had to park a vehicle, sort chairs, or locate items. Before immersion, participants completed the ST and the TMT versions A (TMT-A) and B (TMT-B). The authors identified that the completion time of the second level (Kendall $τ=0.32; P=0.01; P$ value as reported in manuscript) and the number of levels completed in the parking simulator ($τ=0.43; P<0.01; P$ value as reported in manuscript) were correlated with participants’ performance on the ST. In addition, the ST was correlated with seating arrangement metrics, such as time to place the first stool ($τ=0.33; P=0.01; P$ value as reported in manuscript) and number of stools placed ($τ=0.33; P=0.02; P$ value as reported in manuscript), as well as with time to locate the first item in the chemistry laboratory ($τ=0.37; P=0.01; P$ value as reported in manuscript). Correlations between the TMT-A or TMT-B and, for example, the number of completed parking levels ($τ=0.49; P<0.01; P$ value as reported in the manuscript) or the number of items placed in the seating arrangement task in the chemistry laboratory ($τ=0.35; P=0.01; P$ value as reported in the manuscript) were reported. However, reporting was limited to significant correlations only, and no a priori expectation of how performances on the VR and validation tasks were correlated was indicated in the study.

Miskowiak et al [63] assessed executive functions by letting participants complete the TMT-B. One Touch Stockings of Cambridge mean choices to correct, and verbal fluency test versions S and D. The performance on these tests was compared with participants’ performance on a cooking task in VR. The authors hypothesized that the number of cooking tasks that were correctly placed on a to-do list and the latency to solve the task would be VR-equivalent measures of EF. The authors found that VR performance was correlated ($r_{12}=0.26; P=0.004$) with EF, which consisted of a correlation between the average performance on the VR subtasks and the average performance on the validation tasks. The correlations between the individual performances on the VR and validation tasks were not reported in the manuscript.

Pallavicini et al [64] had participants play the Audioshield dance game, which the authors hypothesized could be closely related to EF constructs such as inhibition and working memory. However, the authors correlated participants’ performance on the Audioshield game with their performance on the TMT-A and TMT-B, which measure psychomotor speed (TMT-A) and mental flexibility (TMT-B). Nonetheless, the results showed that TMT performance was negatively correlated with Audioshield performance metrics.

Porffy et al [65] had participants complete VStore, where the 2 tasks measured EF, namely the “Find” task and the “Coffee” task. Specifically, participants had to find 12 items from a list they had previously memorized. In addition, participants had to order a hot drink from the coffee shop after finding, bagging, and paying for the 12 remembered items they had found in the store. Notably, the authors indicated that the 2 VR tasks also tapped into navigation (ie, “Find” task) and processing speed (ie, “Coffee” task). Furthermore, the Groton Maze Learning Test from Cogstate, which the participants completed before the VR task, was used to evaluate general EF. Nonetheless, through their regression analysis, the authors identified that the Groton Maze Learning Test was not a predictor for the “Find” task ($B=0.024; SE 0.029; P=0.11; P$ value as reported in the manuscript) or the “Coffee” task ($B=0.003; SE 0.051; P=0.96; P$ value as reported in the manuscript).

Tan et al [66] had 100 participants complete 13 tasks in a virtual environment that were designed to measure 6 cognitive domains, such as EF and complex attention. Although differences in performance on VR tasks related to EF between age groups were found, no comparison was made with a traditional neuropsychological assessment of EF or any subconstructs of EF.

Tsai et al [67] immersed 2 participant groups in a virtual shopping environment: one group with mild cognitive impairment (MCI) and one control group. The VR tasks assessed participants’ memory, EF, and calculation by having them memorize a shopping list, search for the listed items in the shop, and subsequently pay for them. The authors trained machine learning models on features extracted from the VR tasks to predict whether participants had MCI or were healthy controls, which was achieved with high accuracy. Nonetheless, no neuropsychological assessment of EF was reported as a validation for the VR tasks.

**Targeted Constructs**

The following subsections elaborate on the EF constructs and subconstructs addressed in the studies under review. A range of correlation coefficients were reported in these papers; however, because of the lack of uniformity in results reporting, these coefficients were omitted from the current synthesis. Typically, the papers reported only significant correlations between metrics without presenting all potential correlations. Furthermore, only 16% (3/19) of the studies specified an $α$ level (ie, .05), with another 16% (3/19) of the studies indicating statistical significance at a $P$ value of $≤0.05$. A total of 21% (4/19) of the studies did not indicate an $α$ level but mentioned applying corrections for multiple comparisons, yet they did not detail the adjusted $α$ level. In total, 5% (1/19) of the studies adopted Bayesian statistics using a Bayesian factor of >10 for statistical inference. Nonetheless, in the reviewed studies, it was not consistently clarified which VR tasks were validated against traditional tasks, hindering the construct validity of the various EF components. Consequently, drawing consistent conclusions on how EF constructs of subconstructs were evaluated was not feasible without inferring the nature of the tests and assessment paradigms.

**Core Executive Functions**

**Inhibition**

Of the 3 “core” executive functions, 37% (7/19) of the studies included in our review investigated inhibitory control, interference control, or impulsivity either singly or combined. **Table 2** details the respective validation tasks and target constructs of each of these studies. For example, Chiocchi Giglioli et al [68] presented participants with 6 standardized tasks, 3 of
which assessed inhibition (Table 2), before administering a serious game in which participants were required to perform tasks in outer space. In total, 10 of the 36 possible correlations between measures for the standardized tasks and the serious game tasks were reported as statistically significant and ranged from weak ($0.20 < r < 0.39$; relative $P$ values indicated in the manuscript, eg, $P<.05$) to strong ($0.60 < r < 0.79$; relative $P$ values indicated in the manuscript). For example, the latency metric of the dot-probe task (DPT) correlated positively ($0.35 < r < 0.54$; relative $P$ values indicated) with the latency metric of the 3 VR tasks aimed at measuring inhibition, whereas no correlations were reported between the correct answer metric of the DPT and the correct answer metric of the 3 VR tasks aimed at measuring inhibition. None of the metrics from the ST correlated with those of the VR task (requiring participants to fight aliens); however, the correct answer and latency metrics of the ST correlated with those of the VR task (requiring participants to repair a valve).

Table 2. The validation tasks, authors, and total number of studies examining each construct.

<table>
<thead>
<tr>
<th>VR$^a$ target construct and validation task</th>
<th>Validation</th>
<th>Authors</th>
<th>Studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibition or Inhibitory control</td>
<td></td>
<td></td>
<td>6 (32)</td>
</tr>
<tr>
<td>• DPT$^b$</td>
<td>Implicit</td>
<td>Chicchi Giglioli et al [69]</td>
<td></td>
</tr>
<tr>
<td>• GNG$^c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ST$^d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• DPT</td>
<td>Explicit</td>
<td>Chicchi Giglioli et al [68]</td>
<td></td>
</tr>
<tr>
<td>• GNG</td>
<td>Implicit</td>
<td>Marín-Morales et al [70]</td>
<td></td>
</tr>
<tr>
<td>• ST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• CPT$^e$</td>
<td>Implicit</td>
<td>Voinescu et al [71]$^f$</td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A$^g$</td>
<td>Parsons and Carlew [72]</td>
<td></td>
</tr>
<tr>
<td>• ST</td>
<td>Implicit</td>
<td>Parsons and Barnett [73]</td>
<td></td>
</tr>
<tr>
<td>Interference control</td>
<td></td>
<td></td>
<td>3 (16)</td>
</tr>
<tr>
<td>• ST</td>
<td>Implicit</td>
<td>Marín-Morales et al [70]$^h$</td>
<td></td>
</tr>
<tr>
<td>• The CW-IT$^i$ from the D-KEFS$^j$</td>
<td>Implicit</td>
<td>Parsons and Carlew [72]</td>
<td></td>
</tr>
<tr>
<td>• Automated neuropsychological assessment metrics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• CW-IT from the D-KEFS</td>
<td>Implicit</td>
<td>Parsons and Barnett [73]</td>
<td></td>
</tr>
<tr>
<td>Impulsivity</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Chicchi Giglioli et al [68]</td>
<td></td>
</tr>
</tbody>
</table>

$^a$VR: virtual reality.
$^b$DPT: dot-probe task.
$^c$GNG: Go/No-Go.
$^d$ST: Stroop test.
$^e$CPT: continuous performance test.
$^f$Some traditional tasks listed were included for divergent validity and, therefore, have been omitted from this table.
$^g$N/A: not applicable.
$^h$The VR task involved 42 VR mini-games that assessed various cognitive constructs. A total of 4 mini-games and their target constructs were documented and included in this table; however, the comparative assessments were not provided, and an extensive list of all 42 mini-games was not provided.
$^i$CW-IT: Color-Word Interference Test.
$^j$D-KEFS: Delis-Kaplan Executive Function System.

Similarly, Chicchi Giglioli et al [69] immersed participants in a virtual kitchen in which they had to cook different types of food. The activities were grouped into 4 subtasks of incremental difficulty where, in the third level, inhibition was assessed by determining whether the right dressing was added using a Go/No-Go (GNG)–type paradigm. The authors stated that the DPT, GNG, and ST were used as standard tasks to assess inhibition. The unspecified metric of “correct dressing” was shown to correlate well ($r=0.527; P<.01$; relative $P$ value indicated in the manuscript) with the correct answer metric of
the ST in one group, whereas in the second group, a moderate negative correlation ($r = -0.486; P \leq 0.05$; relative $P$ value indicated in the manuscript) was found between the execution time of the Tower of London task and the correct dressing metric. However, no other correlations between the VR task metric and those of the traditional assessments of inhibition were reported.

Marín-Morales et al [70] had participants complete neuropsychological assessments, including the GNG task, as well as 42 mini-games in VR. An undisclosed set of variables from the mini-games was used as predictors for measures of neuropsychological batteries. The mini-game predictor variables were fed into different machine learning algorithms. The authors highlighted that games related to inhibition produced worse results compared with other games but did not report any results on inhibition. The authors did find that mini-game features of planning and attention could predict GNG hit proportions and mean time with 80% and 94% accuracy, respectively.

Parsons and Carlew [72] had participants perform the ST in a virtual classroom as well as complete a computerized and paper-and-pencil version of the task. The authors found that participants’ performance was lower for color naming and word reading in the VR paradigm than in the paper-and-pencil version but interference performance was better in the VR paradigm than in the paper-and-pencil version. Similarly, Parsons and Barnett [73] had participants perform the ST in a virtual apartment as well as complete a computerized and paper-and-pencil version of the task. Here, the authors found that participants were more accurate in the ST in the virtual apartment version than in the VR paradigm.

Table 3. The validation tasks, authors, and total number of studies targeting working memory.

<table>
<thead>
<tr>
<th>VR target construct and validation task</th>
<th>Validation</th>
<th>Authors</th>
<th>Studies, n (%),</th>
<th>n (%),</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• WAIS-IVb</td>
<td>Implicit</td>
<td>Marín-Morales et al [70]c</td>
<td>4 (21)</td>
<td></td>
</tr>
<tr>
<td>• The Working Memory Index (Digit Span and Arithmetic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• WAIS-IIIbLNSg</td>
<td>Explicit</td>
<td>Miskowiak et al [63]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• CANTABf</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• 1-back and 2-back test (Cogstate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/Ah</td>
<td>Robitaille et al [74]i</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aVR: virtual reality.
bWAIS-IV: Wechsler Adult Intelligence Scale–IV.
cThe VR task involved 42 VR mini-games that assessed various cognitive constructs. In total, 4 mini-games and their target constructs were documented and included in this table; however, the comparative assessments were not provided, and an extensive list of all 42 mini-games was not provided.
dWAIS-III: Wechsler Adult Intelligence Scale–III.
eLNS: Letter-Number Sequencing.
fSWM: Spatial Working Memory.
gCANTAB: Cambridge Neuropsychological Test Automated Battery.
hN/A: not applicable.
iRobitaille et al [74] used a VR paradigm with avatars to trial a dual-task walking protocol.

Porffy et al [65] asked participants to operate a virtual store in which the working memory component was assessed at the “Pay” step, where participants had to select and pay for their items at a self-checkout machine providing the exact amount.
The authors specified that the reaction time on the 1-back task and the accuracy of performance on the 2-back task were metrics from traditional tasks used to assess working memory. Using linear regression, the authors found that performance on the 2-back task was negatively associated ($B=-0.085$; SE 0.042; $P=0.047$) with participants’ performance on the “Pay” step.

Robitaille et al [74] assessed working memory during their simultaneous cognitive tasks, in which participants had to both recognize faces in windows that had been previously declared as “hostile” or “nonhostile” and complete a navigation task. However, no correlations between the traditional and VR tasks were reported.

### Cognitive Flexibility

One study by Chicchi Giglioli et al [68] investigated cognitive flexibility (termed “cognitive shifting” in the paper) through 3 VR tasks. The authors specified that the TMT was used as a traditional task to assess cognitive flexibility as a comparator for the first VR task (CF1, cultivating food) and the Wisconsin Card Sorting Test was used as a traditional task to evaluate cognitive flexibility as a comparator for the other 2 VR tasks.

In another study, Chicchi Giglioli et al [69] used a VR paradigm based on an outer-space environment. The paradigm contained 8 tasks, one of which assessed planning ability (task 7). The authors stated that the Tower of London task was the traditional assessment tool used to evaluate planning and explained that the total score, initial time, and execution time of the VR task were the outcome metrics. Moderate positive correlations were found between the execution time of the VR task and of the Tower of London task ($r=0.372$; $P<.05$). Furthermore, the VR task correlated with some metrics of other traditional assessments used to assess planning ability, although these were not specified a priori.

Both the studies by Kourtesis et al [76] and Kourtesis and MacPherson [75] used the same VR environment based on a variety of everyday tasks. One task assessing planning ability (CF2, growing plants, and CF3, fueling a turbine). The total time metric of the first VR task correlated positively with the total time of the TMT-B ($r=0.396$; $P<.01$; $P$ value as reported in the manuscript), and multiple metrics of VR tasks 2 and 3 correlated with the performance metrics of the Wisconsin Card Sorting Test.

### Higher-Order Executive Functions: Planning

In total, 26% (5/19) of the studies [62,68,69,75,76] identified planning as a target construct in their VR paradigms. Table 4 details the respective validation tasks and target constructs of each of these studies. The VR environment created by Chicchi Giglioli et al [69] used a cooking task with 4 levels of difficulty. In the 3 more difficult levels, planning was required to complete the tasks as 2 burners were used. There was no clearly specified metric for the VR task that was used to evaluate planning, but the authors specified that the Tower of London task was used as a traditional assessment to evaluate planning. A variety of VR task metrics, such as total time to complete a difficulty level, were shown to correlate with various Tower of London task metrics.

### Table 4. The validation tasks, authors, and total number of studies targeting planning.

<table>
<thead>
<tr>
<th>VR² target construct and validation task</th>
<th>Validation</th>
<th>Authors</th>
<th>Studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td></td>
<td></td>
<td>5 (26)</td>
</tr>
<tr>
<td>TOL-DX b</td>
<td>Implicit</td>
<td>Chicchi Giglioli et al [69]</td>
<td></td>
</tr>
<tr>
<td>TOL c</td>
<td>Explicit</td>
<td>Chicchi Giglioli et al [68]</td>
<td></td>
</tr>
<tr>
<td>None specifically reported</td>
<td>N/A d</td>
<td>Davison et al [62] e</td>
<td></td>
</tr>
<tr>
<td>The Key Search task from BADS f</td>
<td>Explicit</td>
<td>Kourtesis et al [76]</td>
<td></td>
</tr>
<tr>
<td>None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td></td>
</tr>
</tbody>
</table>

|                                          |            |         |               |

aVR: virtual reality.
cTOL: Tower of London test.
dN/A: not applicable.
eThe VR task was used to assess executive function. The comparative assessments that validated this assessment were detailed under “executive function” broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.
fBADS: Behavioral Assessment of the Dysexecutive Syndrome.

In another study, Chicchi Giglioli et al [68] used a VR paradigm based on an outer-space environment. The paradigm contained 8 tasks, one of which assessed planning ability (task 7). The authors stated that the Tower of London task was the traditional assessment tool used to evaluate planning and explained that the total score, initial time, and execution time of the VR task were the outcome metrics. Moderate positive correlations were found between the execution time of the VR task and of the Tower of London task ($r=0.463$; $P<.01$; $P$ value as reported in the manuscript) and between the initial time of the VR task and the total time of the Tower of London task ($r=0.372$; $P<.05$). Furthermore, the VR task correlated with some metrics of other traditional assessments used to assess planning ability, although these were not specified a priori.

Both the studies by Kourtesis et al [76] and Kourtesis and MacPherson [75] used the same VR environment based on a variety of everyday tasks. One task assessing planning ability required participants to draw their route around the city (eg, visiting the bakery, supermarket, and library and returning home) on a 3D board. Kourtesis et al [76] explained that the Key Search Task from the Behavioral Assessment of the Dysexecutive Syndrome was used as a traditional measure to assess planning and found a strong positive correlation between the traditional and VR tasks ($r=0.80$; Bayes factor=$4.65 \times 10^5$). Furthermore, Kourtesis and MacPherson [75] noted in their results that planning explained a substantial 12% ($P=.03$) of the variance in time-based prospective memory, which was required in 10 of 17 tasks.

Davison et al [62] assessed planning ability using a task involving the arrangement of a table and a chair. However, they did not explicitly mention the traditional task that was used to evaluate planning. Various correlations between the performance metrics of the VR task and the traditional task were reported.
For example, the performance on the Stroop Color and Word Test was negatively correlated with the time participants took to place a blue chair in the seating arrangement task (Kendall \( \tau = -0.39; P = .01; P \) value as reported in the manuscript).

**Other Domains**

Several studies (14/19, 74%) examined domains of functioning that did not align with the EF definition used in this review. Broadly, these domains fell under the categories of memory, attention, processing, task performance, and a variety of other uncategorized subconstructs. As the literature [1,2,4,6] does not relate these broad domains to EF, they are not discussed further but are presented in Tables 5-6.
Table 5. The validation tasks, authors, and total number of studies targeting constructs classified as uncategorized.

<table>
<thead>
<tr>
<th>VR&lt;a target construct and validation task</th>
<th>Validation</th>
<th>Authors</th>
<th>Studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory (general)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· None specifically reported</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Tsai et al [67]</td>
<td>11 (58)</td>
</tr>
<tr>
<td><strong>Verbal memory and verbal learning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· RAVLT&lt;sup&gt;c&lt;/sup&gt; subtests: total, immediate recall, delayed recall, and recognition</td>
<td>Explicit</td>
<td>Miskowiak et al [63]</td>
<td>2 (11)</td>
</tr>
<tr>
<td>· International Shopping List Test (Cogstate; verbal learning)</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td>2 (11)</td>
</tr>
<tr>
<td><strong>Prospective memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· None specifically reported</td>
<td>N/A</td>
<td>Banville et al [61]&lt;sup&gt;d&lt;/sup&gt;</td>
<td>4 (21)</td>
</tr>
<tr>
<td>· CAMPROMPT&lt;sup&gt;e&lt;/sup&gt; [79]</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;f&lt;/sup&gt;</td>
<td>4 (21)</td>
</tr>
<tr>
<td>· None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td>4 (21)</td>
</tr>
<tr>
<td>· CVLT-II&lt;sup&gt;g&lt;/sup&gt; [80]</td>
<td>Implicit</td>
<td>Parsons and McMahan [53]</td>
<td>4 (21)</td>
</tr>
<tr>
<td><strong>Episodic memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· RBMT-III&lt;sup&gt;h&lt;/sup&gt; [81]</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;f&lt;/sup&gt;</td>
<td>3 (16)</td>
</tr>
<tr>
<td>· CVLT-II</td>
<td>Implicit</td>
<td>Parsons and McMahan [53]</td>
<td>3 (16)</td>
</tr>
<tr>
<td><strong>Immediate recognition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· RBMT-III [81]</td>
<td>Explicit</td>
<td>Kourtesis et al [76]</td>
<td>2 (11)</td>
</tr>
<tr>
<td>· None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td>2 (11)</td>
</tr>
<tr>
<td><strong>Delayed recognition</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· RBMT-III [81]</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;f&lt;/sup&gt;</td>
<td>2 (11)</td>
</tr>
<tr>
<td>· None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td>2 (11)</td>
</tr>
<tr>
<td><strong>Attention</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>General attention</strong></td>
<td></td>
<td></td>
<td>13 (68)</td>
</tr>
<tr>
<td>· DPT&lt;sup&gt;i&lt;/sup&gt;</td>
<td>Implicit</td>
<td>Chicchi Giglioli et al [69]</td>
<td>4 (21)</td>
</tr>
<tr>
<td>· GNG&lt;sup&gt;j&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· ST&lt;sup&gt;k&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· DPT</td>
<td>Explicit</td>
<td>Chicchi Giglioli et al [68]</td>
<td>4 (21)</td>
</tr>
<tr>
<td>· GNG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· ST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· TMT-A&lt;sup&gt;l&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· TMT-B&lt;sup&gt;m&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· DPT—selective attention</td>
<td>Implicit</td>
<td>Marín-Morales et al [70]&lt;sup&gt;o&lt;/sup&gt;</td>
<td>2 (11)</td>
</tr>
<tr>
<td>· GNG—sustained attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· ST—selective attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· TMT&lt;sup&gt;2&lt;/sup&gt;—visual attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>· RVP&lt;sup&gt;8&lt;/sup&gt; CANTAB&lt;sup&gt;9&lt;/sup&gt; (accuracy and latency)</td>
<td>Explicit</td>
<td>Miskowiak et al [63]</td>
<td>2 (11)</td>
</tr>
<tr>
<td>· RBANS-DS&lt;sup&gt;7&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Divided attention</strong></td>
<td></td>
<td></td>
<td>2 (11)</td>
</tr>
<tr>
<td>VR&lt;sup&gt;a&lt;/sup&gt; target construct and validation task</td>
<td>Validation</td>
<td>Authors</td>
<td>Studies, n (%)</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>------------</td>
<td>---------</td>
<td>----------------</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Robitaille et al [74]&lt;sup&gt;4&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• CTT-B&lt;sup&gt;1&lt;/sup&gt; [75,82]</td>
<td>Explicit</td>
<td>Wilf et al [54]</td>
<td></td>
</tr>
<tr>
<td><strong>Complex attention</strong></td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Tan et al [66]</td>
<td></td>
</tr>
<tr>
<td><strong>Selective visual attention</strong></td>
<td></td>
<td></td>
<td>2 (11)</td>
</tr>
<tr>
<td>• The map task from the Test of Everyday Attention</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;6&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td></td>
</tr>
<tr>
<td><strong>Selective auditory attention</strong></td>
<td></td>
<td></td>
<td>2 (11)</td>
</tr>
<tr>
<td>• The Elevator Counting With Distraction task of the Test of Everyday Attention</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;6&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td></td>
</tr>
<tr>
<td><strong>Sustained visual attention</strong></td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• CTT-A&lt;sup&gt;2&lt;/sup&gt; [82]</td>
<td>Explicit</td>
<td>Wilf et al [54]</td>
<td></td>
</tr>
<tr>
<td><strong>Visuospatial attention</strong></td>
<td></td>
<td></td>
<td>2 (11)</td>
</tr>
<tr>
<td>• The Ruff 2 and 7 Selective Attention Test</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;6&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td></td>
</tr>
<tr>
<td><strong>Sustained attention</strong></td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• CPT&lt;sup&gt;4&lt;/sup&gt; [83]</td>
<td>Implicit</td>
<td>Voinescu et al [71]</td>
<td></td>
</tr>
<tr>
<td><strong>Processing</strong></td>
<td></td>
<td></td>
<td>3 (16)</td>
</tr>
<tr>
<td><strong>Processing speed</strong></td>
<td></td>
<td></td>
<td>3 (16)</td>
</tr>
<tr>
<td>• WAIS-IV&lt;sup&gt;o&lt;/sup&gt; Processing Speed Index (symbol search and coding)</td>
<td>Implicit</td>
<td>Marín-Morales et al [70]&lt;sup&gt;o&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• RBANS-CT&lt;sup&gt;x&lt;/sup&gt;</td>
<td>Explicit</td>
<td>Miskowiak et al [63]</td>
<td></td>
</tr>
<tr>
<td>• TMT-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Detection task (Cogstate)</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td></td>
</tr>
<tr>
<td><strong>Task performance</strong></td>
<td></td>
<td></td>
<td>4 (21)</td>
</tr>
<tr>
<td><strong>Dual task</strong></td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• TMT-A</td>
<td>Implicit</td>
<td>Chicchi Giglioli et al [69]</td>
<td></td>
</tr>
<tr>
<td>• TMT-B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multitask</strong></td>
<td></td>
<td></td>
<td>3 (16)</td>
</tr>
<tr>
<td>VR&lt;sup&gt;a&lt;/sup&gt; target construct and validation task</td>
<td>Validation</td>
<td>Authors</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>------------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>• Modified version of the SET&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Implicit</td>
<td>Banville et al [61]&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• CTT&lt;sup&gt;c&lt;/sup&gt; [82]</td>
<td>Explicit</td>
<td>Kourtesis et al [76]&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Kourtesis and MacPherson [75]</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>VR: virtual reality.
<sup>b</sup>N/A: not applicable.
<sup>c</sup>RAVLT: Rey Auditory Verbal Learning Test.
<sup>d</sup>The VR task was predominantly a sorting task for executive function assessment. The comparative assessments that validated this assessment were detailed under “executive function” broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.
<sup>e</sup>CAMPROMPT: Cambridge Prospective Memory Test.
<sup>f</sup>Kourtesis et al [76] explicitly broke episodic memory down into immediate and delayed recognition. However, we gathered these two constructs under episodic memory.
<sup>g</sup>CVLT-II: California Verbal Learning Test–Second Edition.
<sup>h</sup>RBMT-III: Rivermead Behavioral Memory Test–Third Edition.
<sup>i</sup>DPT: dot-probe task.
<sup>j</sup>GNG: Go/No-Go.
<sup>k</sup>ST: Stroop test.
<sup>l</sup>TMT-A: Trail-Making Test version A.
<sup>m</sup>TMT-B: Trail-Making Test version B.
<sup>n</sup>TMT: Trail-Making Test.
<sup>o</sup>The VR task involved 42 VR mini-games that assessed various cognitive constructs. In total, 4 mini-games and their target constructs were documented and included in this table; however, the comparative assessments were not provided, and an extensive list of all 42 mini-games was not provided.
<sup>p</sup>RVP: Rapid Visual Information Processing.
<sup>q</sup>CANTAB: Cambridge Neuropsychological Test Automated Battery.
<sup>r</sup>RBANS-DS: Repeatable Battery for the Assessment of Neuropsychological Status–Digit Span.
<sup>s</sup>Robitaille et al [74] used a VR paradigm with avatars to trial a dual-task walking protocol.
<sup>t</sup>CTT-B: Color Trails Test B.
<sup>u</sup>CTT-A: Color Trails Test A.
<sup>v</sup>CPT: continuous performance test.
<sup>w</sup>WAIS-IV: Wechsler Adult Intelligence Scale–IV.
<sup>x</sup>RBANS-CT: Repeatable Battery for the Assessment of Neuropsychological Status–Coding Test.
<sup>y</sup>SET: Six Elements Test.
<sup>z</sup>CTT: Color Trails Test.
Table 6. The validation tasks, authors, and total number of studies targeting constructs classified as uncategorized.

<table>
<thead>
<tr>
<th>VR(^a) target construct and validation task</th>
<th>Validation</th>
<th>Authors</th>
<th>Studies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncategorized(b)</td>
<td></td>
<td></td>
<td>12 (63)</td>
</tr>
<tr>
<td>Visual perception</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A(^c)</td>
<td>Marín-Morales et al [70](^d)</td>
<td></td>
</tr>
<tr>
<td>Verbal learning</td>
<td></td>
<td></td>
<td>2 (11)</td>
</tr>
<tr>
<td>• RAVLT(^e) subtests: total, immediate recall, delayed recall, and recognition</td>
<td>Explicit</td>
<td>Miskowiak et al [63]</td>
<td></td>
</tr>
<tr>
<td>• International Shopping List Test (Cogstate)</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td></td>
</tr>
<tr>
<td>Navigation</td>
<td></td>
<td></td>
<td>2 (11)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Porffy et al [65]</td>
<td></td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Robitaille et al [74]</td>
<td></td>
</tr>
<tr>
<td>Associate learning</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• Continuous Paired Associate Learning Test (Cogstate)</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td></td>
</tr>
<tr>
<td>Pattern recognition</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• Continuous Paired Associate Learning Test (Cogstate)</td>
<td>Implicit</td>
<td>Porffy et al [65]</td>
<td></td>
</tr>
<tr>
<td>Perceptual motor</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Tan et al [66]</td>
<td></td>
</tr>
<tr>
<td>Social cognition</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Tan et al [66]</td>
<td></td>
</tr>
<tr>
<td>Learning and memory</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Tan et al [66]</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Tan et al [66]</td>
<td></td>
</tr>
<tr>
<td>Calculation</td>
<td></td>
<td></td>
<td>1 (5)</td>
</tr>
<tr>
<td>• None specifically reported</td>
<td>N/A</td>
<td>Tsai et al [67]</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)VR: virtual reality.
\(^b\)Williams et al [55] replicated the Wisconsin Card Sorting Test and multitasking task but did not explicitly state the cognitive constructs that the VR task was assessing. For this reason, the paper has not been assigned a target construct.
\(^c\)N/A: not applicable.
\(^d\)The VR task involved 42 VR mini-games that assessed various cognitive constructs. In total, 4 mini-games and their target constructs were documented and included in this table; however, the comparative assessments were not provided, and an extensive list of all 42 mini-games was not provided.
\(^e\)RAVLT: Rey Auditory Verbal Learning Test.

Validity and Reliability

Tables 1-6 show details of the current validated comparator tasks against the novel VR tasks if they were explicitly provided by the authors. Where “None specifically reported” is stated, the authors of each paper did not identify or indicate a direct comparator. All but 2 studies (17/19, 89%) [72,73] set out to assess multiple constructs. In some cases, the subconstructs that were assessed were individually validated against existing validated tasks. In other cases, a suite of existing validated tasks was included in the analysis for correlation against a variety of subconstructs being assessed using the VR battery. In these cases, there was no validation at the construct level identified a priori. In 16% (3/19) of the studies, there was no reported validation of the VR paradigm.

Notably, only one study used real-life validation criteria in addition to construct-driven tests to present a validation of their VR paradigm. Specifically, Miskowiak et al [63] functionally assessed participants using the Functioning Assessment Short Test (FAST) and the brief University of California, San Diego, Performance-Based Skills Assessment (UPSA-B). Participants’
scores on these assessments were correlated with their performance on the test domains of the VR paradigm, called cognition assessment in VR (CAVIR). The authors identified that participants’ performance on the FAST was negatively associated ($r=0.17$; $P<0.30$; no exact or relative $P$ values reported) with CAVIR test domains such as processing speed and working memory, whereas participants’ performance on the UPSA-B was positively associated with the CAVIR test working memory ($r=0.40$; $P$ value not exactly or relatively reported) and cognition composite ($r=0.44$; $P<0.001$) domains. Moreover, the authors noted that lower global scores on traditional (ie, construct-led) neuropsychological tests were negatively associated with FAST scores ($r=-0.45$; $P<0.001$) and positively associated with UPSA-B scores ($r=0.52$; $P<0.001$), highlighting that lower CAVIR scores were associated with more functional disability, as indicated by the functional and traditional assessment tools.

The reliability of the VR paradigm was only assessed in 5% (1/19) of the studies. This was done by Kourtesis et al [76], who reported good internal reliability (Cronbach $\alpha=0.79$) of their VR Everyday Assessment Lab (EAL) paradigm. However, this global internal consistency report did not provide a reliability estimate of the unique cognitive functions targeted by their VR EAL paradigm. Nonetheless, none of the reviewed studies included a test-retest analysis to highlight the reliability of their VR paradigm.

**Evaluation of User Experience, Cybersickness, Immersion, and Engagement**

An overview of the measures used to evaluate participants’ experiences and well-being can be found in Multimedia Appendix 1 [53-55,61-76]. Of the 19 studies, 5 (26%) included user experience assessments. To measure participants’ virtual presence, experience, and well-being, the studies administered the Igroup Presence Questionnaire [61], Presence Questionnaire [63,71,74], or Slater-Usoh-Steed questionnaire [74]. To measure participants’ discomfort, the studies used the Simulator Sickness Questionnaire [61,71,74] or an adaption of it, the Virtual Reality Sickness Questionnaire [63]. To evaluate the usability of the virtual environment, the studies used the System Usability Scale [71]. To measure participants’ virtual experience and comfort, 11% (2/19) of the studies used the Virtual Reality Neuroscience Questionnaire [76].

Two studies (2/19, 11%) investigated whether system usability, virtual presence, or cybersickness affected participants’ task performance. For example, Porföy et al [65] measured participants’ technical familiarity and found that it explained between 10% and 42% of the variability in participants’ performance on the VStore outcomes “Recall”, “Find”, and “Select”. Conversely, participants’ technical familiarity appeared to influence their performance on VStore. Kourtesis et al [76] used questionnaires to evaluate the quality of the VR paradigm, participants’ gaming experience, and the realism (verisimilitude) and pleasantness of the VR paradigm. The authors identified no relationship between VR experience, gaming experience, and performance on the VR EAL tasks.

Some papers (4/19, 21%) reported on cybersickness, presence, or usability scores but did not report an analysis of the relationship between task performance and measures evaluating the VR paradigm. For example, Banville et al [61] recorded participants’ sickness and virtual presence but did not report any test evaluating whether sickness or presence affected task performance. Similarly, Voinescu et al [71] obtained system usability ratings from participants; however, no test was reported wherein the effect of usability on task performance was assessed. Finally, Chicchi Giglioli et al [68] recorded participants’ use of technology but did not report an analysis between technology use and task performance.

Finally, some studies (2/19, 10%) evaluated participants’ experiences post hoc, although it was not disclosed whether any validated scales were used. For example, Davison et al [62] measured participants’ enjoyment of the VR tasks and their preference for either the VR tasks or the pencil-and-paper tasks. The authors found that younger participants rather than older ones preferred VR tasks over pencil-and-paper tasks. In addition, 11 out of 40 participants reported having experienced a mild degree of motion sickness. However, 58% (11/19) of the papers did not disclose any information about user experiences.

**Discussion**

**Overview**

The purpose of this review was to investigate the development and validation of VR assessment tools for EF. Specifically, we examined the components of EF that were assessed using VR, their validation processes, and whether immersion and cybersickness assessments were used. Although research in this domain is proliferating, the results of this review suggest that the process of development and validation varies considerably between studies.

**Components of EF Assessed Using VR Paradigms**

**Overview**

The terminology used in the papers to describe EF constructs was inconsistent. For example, the most popular construct set assessed using VR comprised the inhibition processes. “Inhibitory control” encompasses the inhibition of goal-irrelevant stimuli, cognitions, and behavioral responses [6,84]. In total, two of the key components of inhibitory control are response inhibition and attentional inhibition [85]. Response inhibition was also termed “inhibition control,” “prepotent response inhibition,” and “motor inhibition,” whereas attentional inhibition was also termed “control of interference,” “interference control,” and “external interference control.” Although these terms are used in the literature [85], its readability and synthesis would be improved through agreement on a particular term for the same construct. In the same way, several studies (7/19, 37%) examined “EF” broadly without specifically detailing its components. In these studies, EF was validated using different measurement tools, suggesting that, across studies, EF was defined and used differently in each VR paradigm. As the constructs that these paradigms aimed to assess were not explicitly detailed, this poses a risk of hampering researchers wishing to build upon previous findings.
Furthermore, there was a broad range of constructs that were not commonly considered as EF domains but were reported as components of EF, making it difficult for future research to replicate the findings of undefined target constructs. For example, several papers (14/19, 74%) reported on verbal learning [63], associate learning pattern recognition [65], perceptual motor, social cognition, language [66], and calculation [67]. Although many of these components rely on EF domains or underpin those domains, they exist at various levels of abstraction. Thus, although the reviewed studies investigated components at different levels and used different languages, it is possible that they overlapped. For example, “organization” may be an umbrella term for a range of EF domains, each of which uses different terminology for the same concept, such as “cognitive flexibility,” “flexible updating,” and “working memory.” Although “organization” is not measured as a higher-order version of the subcomponents, it is difficult for the research that has examined cognitive flexibility and working memory to be extended. Thus, 2 studies assessing the same construct are not able to build on each other’s progress.

**Recommendation: Establish a Coherent and Consistent Framework for EF Terminology**

The Research Domain Criteria (RDoC) framework developed by the National Institute of Mental Health could serve as a framework to address this recommendation. The RDoC was originally created to consolidate the research conducted in various fields of mental health [86]. The framework categorizes cognition into 6 domains and encourages the investigation of these domains via different classes of variables, such as behavioral, physiological, and self-report data. This framework encourages a common language and organizes findings in such a way that researchers can identify gaps or discrepancies in the literature and contribute to the ongoing development of the field. This framework indicates the potential benefits of using a common language for research, although it is not necessarily the only option in this field. Alternatively, researchers could engage in a Delphi study to generate expert-informed consensus on the key constructs of EF that merit investigation using VR paradigms (eg, see the study by Yücel et al [87] for a Delphi study on neuropsychological assessment for addiction). Nonetheless, the emerging area of VR development for neuropsychological assessments would benefit from using the RDoC framework to coordinate the research process.

**Validation of VR for EF**

**Overview**

Overall, there was limited reporting on the constructs that were assessed using VR paradigms and the associated validation outcome measures. In some papers, there was inadequate reporting of the constructs that the VR paradigm was intended to assess. In others, the same construct was assessed using a variety of traditional tasks. Furthermore, some VR paradigms were intended to replicate real life yet were validated against traditional tasks, none of which assessed ecological validity. In some studies, the correlations between the VR paradigm and the traditional tasks were incomplete. Finally, sample sizes varied considerably between studies, also affecting the evaluation of their psychometrics. These points are expanded upon in this section.

Several studies (5/19, 26%) examined EF as a broad category and then validated the paradigm against a variety of traditional tasks. However, some studies (3/19, 16%) detailed limited (or no) reporting of which aspect of the VR paradigm each traditional task was intended to validate. That is, no details were provided regarding which traditional task outcome measure corresponded to each component of EF within the VR paradigm. Traditional tasks, which often target one construct, were then correlated against seemingly all outcomes of the VR paradigm. Although this practice may be beneficial during the exploratory phase of VR paradigm development, failure to correct for multiple comparisons may provide misleading results whereby a correlation is found between two constructs incidentally. Conversely, some traditional tasks assessed multiple constructs, which poses a slightly different challenge. For example, if the VR paradigm broadly assessed EF but was validated against the ST, it was then unclear whether the VR paradigm aimed to assess processing speed, attention, inhibitory control, or interference control as the ST could be used to measure all four. Similarly, when these studies used multiple traditional assessments, the reader was expected to presume the target constructs of the VR paradigm as this was not clearly outlined. Poorly defined target constructs and failure to specify which traditional task validates which aspect of the VR task produces a literature that is difficult to interpret. Moreover, this general lack of clarity means that future researchers are more likely to invent a new paradigm rather than adopt or extend existing paradigms, creating inefficiency and hampering progress in the field.

Various standardized tasks were used to validate target constructs in the VR paradigm. For example, the study by Chicchi Giglioli et al [69] examined attention and inhibition control using the DPT, GNG, and ST. However, Voinescu et al [71] examined inhibition using the CPT paradigm. In addition, Marín-Morales et al [70] assessed inhibition using one mini-game of their VR paradigm. However, they neither provided details of a specific comparator task for validation purposes nor reported the statistical outcomes. Furthermore, the DPT, which is typically used to assess selective attention [88], was used to assess inhibition, although its own psychometric properties have been the subject of controversy [89,90]. Although several traditional tasks purport to measure the same construct (ie, there is not one task for one construct), the lack of consistency between studies makes it difficult to compare VR platforms. Furthermore, the traditional comparator task used to validate the VR paradigm needs to have sound psychometric properties in its own right to assess the respective construct; when two tasks are compared with one another, it is unclear which task may be responsible for discrepancies in the outcome [91]. These points are especially pertinent for studies that rely solely on traditional measures to validate tasks in the absence of other validation techniques.

Although it is promising to see that VR paradigms are being used for ecologically valid assessments, their validation remains a challenge. In the case of traditional tasks, we assume that a single construct can be assessed using a behavioral task and
that the performance on that task is linear with the cognitive construct. In the case of a “function-led” VR task, there is a behavioral task that simulates real-world functioning, which is thought to deteriorate in an EF-declining population. This VR task is not a direct assessment of a target construct—it is a test of a real-world function, such as parking a car. To test convergent validity, the individual would have to park a car in real life and have their performance assessed similarly to that on the VR task and compared. However, when we use traditional measures to validate the “function-led” VR measures, we assume that EF can be reliably measured and the function-led VR task (eg, parking a car) requires the same EF. Thus, those who perform poorly on a traditional EF task are also expected to perform poorly on real-life tasks requiring EF. Critically, if our results do not show this relationship, it could be that the traditional task is a poor test of EF, the function-led assessment is a poor test of EF, or the EF at hand is not related to the functional task (eg, parking a car).

These assumptions place substantial weight on the selection of the traditional task for validating the VR paradigm for predictive validity. Davison et al [62] assessed EF using the ST and TMT. They broadly hypothesized that there would be correlations between the traditional measures and the VR paradigm, which contained tasks that replicated real life, such as car parking, arranging seating, and locating items. In the reported results, the ST and TMT were correlated with all outcome measures of the VR paradigm. For example, performance on the Stroop Color and Word Test was correlated with performance on the second parking simulator task, the number of levels completed on the parking simulator task, and the time taken to place the blue chair in the seating arrangement task. If the ST and TMT are not sufficient validators of the functional task, this may generate misleading results regarding the integrity of the VR paradigm and its ability to sensitively measure EF. Thus, the convergent validity of VR tasks would be better assessed through real-life performance on the same task, such as actually parking in a controlled environment. Although this may seem to be a resource burden to validation, it could provide integral merit to using the paradigm as a proxy for the real-life task thereafter. Alternative options are to assess convergent validity through other forms of real-life functioning (eg, self-care, residence, transportation, and employment) and diagnostic trajectory [49]. Moreover, predictive validators should be carefully chosen to ensure that their target construct aligns with that thought to be required for the function-led assessment.

Nonetheless, for novel task validation, transparent reporting of all results is crucial for advancing future research. Several papers included in this review (4/19, 21%) [61,62,68,69] reported only statistically significant correlations, leaving unanswered questions because of the omission of nonsignificant results. For instance, Chicchi Giglioli et al [69] sought to evaluate inhibition control using the GNG and ST for validation (both are common tasks for assessing inhibition) as well as the DPT yet did not report all correlational data in their results table. Such omissions hinder the comprehensive use or meta-analytic application of these findings. Conversely, Chicchi Giglioli et al [68] provided a detailed comparison between each validation task and its corresponding VR task, including the constructs assessed. However, only significant correlations were reported, some of which were between tasks intended to assess disparate constructs, such as the correlations between the Wisconsin Card Sorting Test (assessing cognitive shifting) and the VR tasks (measuring attention and inhibition control). Although these findings may indicate overlapping constructs in VR tasks, the absence of multiple-comparison correction and a detailed post hoc analysis of these correlations limits the interpretability and applicability of these results.

Finally, it is worth noting that there was significant variation in sample sizes across the studies reviewed. Although it is often accepted that pilot studies or preliminary studies have small sample sizes that often result in underpowered analysis, the utility of the VR paradigms is dependent on sound psychometric properties that require adequate sample sizes and statistical power. As detailed in Multimedia Appendix 1 [53-55,61-76], the sample sizes varied from 12 (6 per group) [74] to 103 (divided into 2 groups) [53]. Although the definition of a “sufficient” sample size may vary between studies and analyses, several of the included VR paradigms would likely require additional validation studies to provide confidence in their psychometric properties.

**Recommendations**

Our recommendations are as follows:

1. Papers should explicitly detail how their VR paradigms are being validated. If a paradigm has multiple components, it is essential to state how each one is being validated. A good example is the paper by Kourtesis et al [76] in this review.

2. If studies aim to validate a VR paradigm for a specific EF construct, they should identify a priori the precise outcome measures of the VR paradigm that are hypothesized to tap into various EF constructs (eg, time to completion and number of errors) and then validate them against the appropriate traditional tasks that also reliably assess those EF constructs.

3. Where appropriate, the VR paradigm’s real-world task should be validated against both traditional task measures and ecologically valid measures. Ecologically valid measures may include carer reports, observation assessments, and activity of daily living assessments.

4. Multiple modes of validation should be used, including measures that provide predictive power [49], and both carer reports of daily functioning and biosensor data should be considered.

5. Papers should report all outcomes of validation data (even those in supplementary materials) to ensure the transparency of the tools’ properties. A concerted effort to increase explicit and transparent reporting would greatly benefit this field.

6. To validate the VR paradigm, the psychometric properties of the traditional task must be appropriate.

7. Studies aiming to evaluate the psychometric properties of their VR paradigm should ensure that they have adequate sample sizes for a powered analysis.
Cybersickness

Overview

Although VR offers several key advantages over traditional tasks, these systems can also produce adverse effects such as cybersickness. In our review, only 21% (4/19) of the studies included an assessment of cybersickness. This is concerning as cybersickness presents a substantial confound for valid VR assessment and has been shown to negatively affect task performance [92,93]. Given that the assessment of EF involves ascertaining a participant’s cognitive abilities, the recording of cybersickness is key to ensuring that common side effects such as dizziness and vertigo do not affect the participants’ ability to perform at their best on the tasks. Without formal evaluation, the degree to which participants’ experiences are altered is unclear. Furthermore, it is unknown at this stage whether cybersickness symptoms affect various client populations differently. For example, it is possible that, although a healthy individual may be able to continue the assessment with minor vertigo, an individual with cognitive impairment may be more affected, resulting in severely affected cognitive outcomes. Thus, caution should be exercised when using VR paradigms to ensure that the potential benefits of engagement and ecological validity are not realized at the cost of the potential negative effects of cybersickness.

Recommendations

Our recommendations are as follows:

1. Future papers should include usability data in the form of cybersickness measurements.
2. Correlations between cybersickness and participants’ task performance could be included as supplementary material that should be accessible to readers, enabling them to better understand how the VR battery is performing.
3. Even when a paradigm has already assessed cybersickness, we encourage future researchers to use the same paradigm to conduct their own cybersickness assessments. This is because it is still unclear whether cybersickness will have different effects on various populations.
4. Clinical researchers and engineers should continue to investigate and report on techniques and technologies that reduce the incidence or severity of cybersickness.

Textbox 1 provides an overview of the recommendations of this review.

<table>
<thead>
<tr>
<th>Validate against multiple forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Examples include carer reports, observation assessments, ecological momentary assessments, activity of daily living assessments, physiological sensors, and in vivo studies.</td>
</tr>
<tr>
<td>- Consider longitudinal tracking of participants to establish predictive utility to initially validate the novel paradigm.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Report a priori how each assessment in the VR paradigm is being validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>- If there are multiple components to one paradigm, state how each element is being validated (a good example is the study by Kourtesis et al [76] in this review); for example: “Task 2a aims to assess inhibitory control and is validated against the traditional stop signal task and go/no-go task.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Report all validation data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Report correlations of all aspects of a task that were identified a priori as validating the VR paradigm. In extending the previous example, show all relevant metrics from task 2a, such as errors, proportion of successful stops, reaction time, and stop signal reaction time against the relevant metrics of both the Stop Signal and Go/No-Go tasks.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Include user experience assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Conduct assessments of immersion, cybersickness, usability, and engagement.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Use a common framework for defining target constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>- The Research Domain Criteria is one option of a framework that can be applied to ensure that terminology used in the field is consistent.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consider adding biosensors</th>
</tr>
</thead>
<tbody>
<tr>
<td>- These provide additional objective data that may inform the VR-based EF assessment.</td>
</tr>
</tbody>
</table>

Limitations

We searched for articles that used the terms “executive functioning,” “higher order cognition,” and “functional assessment” to capture tasks that aimed to broadly assess facets of EF. This search strategy may have missed studies that examined a key construct of EF but did not specifically use the aforementioned terms (eg, used VR to assess inhibitory control alone). In addition, we did not contact the authors of the papers included in this review for further information; however, one of the key outcomes of this review was the amount of information contained in the manuscripts for future studies to extend upon.

Future Directions

The authors posit that the integration of biosensors into a VR system has significant potential. Biosensors such as pupillometry, eye gaze, EEG, and language and grammatical...
characteristic data can be temporally linked to the events occurring in the VR task. For example, pupillometry can offer insights into brain injury prognosis [94] and differentiate between participants with Alzheimer disease and healthy controls [95]. Eye tracking during reading aids Alzheimer disease identification [96], and linguistic attributes (eg, formation and fluency of sentences, syntax, and grammar) distinguish patients with Alzheimer disease from those with MCI [97]. The combination of these biosensor metrics and real-time function-led VR performance could increase the sensitivity of tests, enabling the detection of subtle differences such as between MCI and subjective memory complaints [98]. However, currently, biosignals are rarely evaluated alongside emerging VR paradigms for EF assessment. None of the reviewed studies used biosensors, leaving an untapped potential for VR paradigms to be frontline neuropsychological assessments.

Biosensors could also assist in modulating the cognitive load experienced by participants. Cognitive load is the cumulative working memory resources that an individual requires for a given task [99]. Similar to the gaming industry, VR paradigms could be adaptive and performance driven so that the level of challenge adjusts according to real-time individual responses [100,101]. Modulating the cognitive load adjusts the challenge of a task and enables all participants to encounter similar levels of perceived difficulty for their respective abilities. EEG, pupillometry, and cardiovascular measures are also sensitive to cognitive load capacities [99].

An additional advantage of VR is its ability to facilitate the assessment of spatial navigation. Spatial navigation is a component of cognitive functioning that can be a key factor in detecting early stages of neurodegenerative diseases. However, it cannot be assessed adequately by means of many traditional assessments. Although it is acknowledged that spatial navigation is not a component of EF, the authors of this paper consider it a generally underexamined construct when assessing cognition and general function. For example, spatial navigation is a cognitive marker used to detect early attention deficit [102,103] and offers additional relevant information beyond the traditional neuropsychological tests [103]. The environment could also be systematically manipulated to match the needs of the assessment [104] and tailored to specific populations. However, typically, spatial navigation is assessed using a real-space human analog of the Morris water maze test, which can be difficult to implement under standardized conditions. Computerized versions have been adapted, with findings comparable with those of tests conducted in real space [105], suggesting promise for translating this style of assessment to VR.

Conclusions
VR paradigms assessing EF have great potential to improve upon traditional tests. However, despite their undeniable novelty and potential, their methodological and psychometric properties must be addressed during their development to ensure their validity and reliability. Although there is no shortage of research in this area, the lack of standardized protocols to validate VR-based neuropsychological assessments hinders the progress of this field of research and practice. It is hoped that this study will be the beginning of a larger movement toward systematizing the development and validation of these paradigms.

Acknowledgments
LK received funding support from the Institute for Intelligent Systems Research and Innovation. This funding source had no role in the data analysis, presentation, or interpretation and write-up. MY received funding from government funding bodies such as the National Health and Medical Research Council; Australian Research Council; Australian Defence Science and Technology Group; Department for Industry, Innovation, and Science; and National Institutes of Health (United States), as well as philanthropic donations from the David Winston Turner Endowment Fund and Wilson Foundation; sponsored investigator-initiated trials including Incannex Healthcare Ltd; and payments in relation to court, expert witness, or expert review reports. These funding sources had no role in the data analysis, presentation, or interpretation and write-up. MY also sits on the advisory boards of the Centre for Urban Mental Health, University of Amsterdam; Monash Biomedical Imaging; and Enosis Therapeutics. The authors would also like to acknowledge Scribo Proofreading and Editing for providing their services for this manuscript.

Data Availability
All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Authors’ Contributions
RK contributed to conceptualization (equal), data curation (equal), formal analysis, investigation, methodology, writing—original draft preparation (lead), and writing—review and editing. LK contributed to conceptualization (equal), formal analysis (lead), investigation, writing—original draft preparation, and writing—review and editing. KR contributed to conceptualization (equal), data curation (equal), writing—original draft preparation, and writing—review and editing. MY contributed to writing—review and editing and funding acquisition. LA and DM contributed to writing—review and editing.

Conflicts of Interest
None declared.
References


Abbreviations

CAVIR: cognition assessment in virtual reality
CPT: continuous performance test
DPT: dot-probe task
EAL: Everyday Assessment Lab
EEG: electroencephalography
EF: executive functioning
FAST: Functioning Assessment Short Test
GNG: Go/No-Go
MCI: mild cognitive impairment
MET: Multiple Errands Test
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RDoC: Research Domain Criteria
ST: Stroop test
TMT: Trail-Making Test
TMT-A: Trail-Making Test version A
TMT-B: Trail-Making Test version B
UPSA-B: brief University of California, San Diego, Performance-Based Skills Assessment
VMT: Virtual Multitasking Test
VR: virtual reality
WoS: Web of Science
Review

Digital Gamification Tools to Enhance Vaccine Uptake: Scoping Review

Hina Hakim1, PhD; S Michelle Driedger2, PhD; Dominique Gagnon3, MSc; Julien Chevrier4, MLIS; Geneviève Roch5,6,7, MSc, RN, PhD; Eve Dubé5,8, PhD; Holly O Witteman1,6,7, PhD

1Department of Family and Emergency Medicine, Université Laval, Québec City, QC, Canada
2Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada
3Direction des risques biologiques, Institut national de santé publique du Québec, Québec City, QC, Canada
4Bibliothèque Louise-Lalonde-Lamarre, Polytechnique Montréal, Montréal, QC, Canada
5Faculty of Nursing, Université Laval, Québec City, QC, Canada
6Centre hospitalier universitaire (CHU) de Québec-Université Laval, Université Laval, Québec City, QC, Canada
7VITAM Research Centre for Sustainable Health, Université Laval, Québec City, QC, Canada
8Département d’anthropologie, Université Laval, Québec City, QC, Canada

Corresponding Author:
Hina Hakim, PhD
Department of Family and Emergency Medicine
Université Laval
1050 avenue de la Médecine, Pavillon Ferdinand-Vandry
Québec City, QC, G1V 0A6
Canada
Phone: 1 418 656 2131
Email: hina.hakim.1@ulaval.ca

Abstract

Background: Gamification has been used successfully to promote various desired health behaviors. Previous studies have used gamification to achieve desired health behaviors or facilitate their learning about health.

Objective: In this scoping review, we aimed to describe digital gamified tools that have been implemented or evaluated across various populations to encourage vaccination, as well as any reported effects of identified tools.

Methods: We searched Medline, Embase, CINAHL, the Web of Science Core Collection, the Cochrane Database of Systematic Reviews, the Cochrane Central Register of Controlled Trials, Academic Search Premier, PsycInfo, Global Health, and ERIC for peer-reviewed papers describing digital gamified tools with or without evaluations. We also conducted web searches with Google to identify digital gamified tools lacking associated publications. We consulted 12 experts in the field of gamification and health behavior to identify any papers or tools we might have missed. We extracted data about the target population of the tools, the interventions themselves (eg, type of digital gamified tool platform, type of disease/vaccine, type and design of study), and any effects of evaluated tools, and we synthesized data narratively.

Results: Of 1402 records, we included 28 (2%) peer-reviewed papers and 10 digital gamified tools lacking associated publications. The experts added 1 digital gamified tool that met the inclusion criteria. Our final data set therefore included 28 peer-reviewed papers and 11 digital gamified tools. Of the 28 peer-reviewed papers, 7 (25%) explained the development of the tool, 16 (57%) described evaluation, and 2 (7%) reported both development and evaluation of the tool. The 28 peer-reviewed papers reported on 25 different tools. Of these 25 digital gamified tools, 11 (44%) were web-based tools, 8 (32%) mobile (native mobile or mobile-enabled web) apps, and 6 (24%) virtual reality tools. Overall, tools that were evaluated showed increases in knowledge and intentions to receive vaccines, mixed effects on attitudes, and positive effects on beliefs. We did not observe discernible advantages of one type of digital gamified tool platform, type of disease/vaccine, and type of study), and any effects of evaluated tools, and we synthesized data narratively.

Conclusions: Digital gamified tools appear to have potential for improving vaccine uptake by fostering positive beliefs and increasing vaccine-related knowledge and intentions. Encouraging comparative studies of different features or different types of digital gamified tools could advance the field by identifying features or types of tools that yield more positive effects across various populations.
Introduction

Vaccination is one of the most cost-effective methods of preventing the spread of vaccine-preventable diseases. If vaccination coverage falls below the thresholds that are safe for the prevention of epidemic transmission, the incidence of vaccine-preventable diseases increases [1,2]. For example, measles returned over the past 2 decades, and the incidence of measles in the European Union increased in 2017-2018 [3].

In 2019, prior to the COVID-19 pandemic, the World Health Organization identified vaccine hesitancy (ie, the reluctance or refusal to be vaccinated despite the availability of vaccination services) as 1 of the top 10 threats to worldwide health [4]. Vaccine hesitancy is one of the several reasons some people are un- or undervaccinated [5-9]. Interventions addressing vaccine hesitancy are therefore necessary to promote vaccine acceptance and uptake. As the contributors of vaccine acceptance are diverse, no single intervention will solve this issue [10]. Multicomponent interventions tailored to local barriers to vaccine acceptance and uptake are known to be the most effective [11,12]. Misinformation and conspiracy theories spread online, where extensive antivaccine content is shared [13-15], potentially negatively influencing views about vaccines [16,17]. Efforts have been made to counter vaccine misinformation and mistrust by targeting various groups, such as parents, non–health care workers [18,19], and adolescents [20], and delivering information about the risks and benefits of different types of vaccines, for instance, human papillomavirus (HPV) vaccination [21] and measles, mumps, and rubella vaccines [22,23]. Along with traditional communication strategies, the use of other strategies to inform and educate about immunizations, for example, with digital gamified tools, may help encourage vaccine uptake.

Gamification is defined as the use of game design elements in non-game contexts [24]. It includes several aspects and features, such as fun interfaces, immediate success or feedback, reward systems (levels, point scores, badges), challenges and competitions, team playing, avatars, and quizzes. Previous studies have used gamification to achieve desired health behaviors [25-27] or facilitate their learning about health [28]. Gamification draws on elements from serious games, meaning fully developed digital games used to train and educate players [29,30]. For example, a serious game “Land of Secret Gardens” facilitates conversations about HPV with preteens. In the game, preteens need to protect their bodies with a “potion,” which offers a metaphor for the HPV vaccine [31]. However, serious games and digital gamified tools are distinct but related concepts. Serious games use gaming as a central and primary medium [32]. In contrast, digital gamified tools (eg, apps) or gamified interventions are not complete game experiences but have gaming features, such as rewards systems, scoring of points, or engaging users in different challenges [33]. In this study, we defined digital gamified tools as digital apps with the aforementioned gaming features. Our definition includes serious games that meet the criteria, that is, they must include such gaming features. This scoping review provides insight into the reported effects of digital gamified tools to increase vaccine uptake. Our review built upon existing reviews in the field by including a comprehensive search of both published literature and online tools, as well as an examination of both the characteristics and the reported effects of these digital tools. This review was distinct in that it focused specifically on digital gamified tools and their effects, rather than simply the effectiveness of gamification in general. In doing so, this review aimed to fill a gap in the literature by providing evidence-based answers to the question of whether gamification “works” to increase vaccine uptake. Therefore, the objectives of this scoping review [34] were, first, to review digital gamified tools that have been implemented or evaluated across various populations to encourage vaccine uptake and, second, to describe any reported effects of the identified tools in terms of influence on users’ knowledge or behavior toward vaccination. Our research questions can therefore be summarized as follows:

• What digital gamified tools intended to encourage vaccination exist and have been described in the literature?
• Do these tools demonstrate any effects on knowledge, attitudes, beliefs, and behaviors about vaccination?

Methods

Search Strategy

For peer-reviewed papers, we searched Medline (Ovid), Embase (Ovid), CINAHL (EBSCO), the Web of Science Core Collection, the Cochrane Database of Systematic Reviews (Ovid), the Cochrane Central Register of Controlled Trials (Ovid), Academic Search Premier (EBSCO), PsycInfo (Ovid), Global Health (Ovid), and ERIC (Ovid) with no language or date restrictions. The proposed search terms were, for example, “vaccine,” “vaccination,” “immunization,” “video games,” “gamification,” “application,” and “virtual reality” (see Multimedia Appendix 1 for the full search strategy). The search was conducted on January 26 and 27, 2022.

We also conducted an online Google search on May 5, 2022, for any digital tools with gamified features that deliver informative or educative messages on vaccination. The search terms we used were “vaccination,” “immunization,” “electronic game,” “computer game,” “mobile game,” “interactive game,” and “digital game” (see Multimedia Appendix 1 for the full digital gamified tools; digital game; vaccination; gamification; vaccine uptake; scoping review; review method; vaccine; gamified; COVID-19; COVID; SARS-CoV-2; health behaviour; health behavior; health promotion; behavior change; behaviour change
search strategy). We reviewed the first 30 results for each search, as it is rare for users to click past the third page of 10 search results per page, and therefore, researchers analyzing medical content available on the web often use 30 as a threshold [35-37]. On May 6, 2022, we conducted the same searches in private browsing mode to ascertain whether our results had been affected by a “filter bubble” [38], that is, the way Google search results are adapted to one’s previous browsing activity. Our search strategy was constructed and reviewed by 2 librarians. Following the librarians’ advice, we expanded our search strategy to include ERIC and Global Health databases.

**Study Selection and Screening Process**

We used PICO (Population, Intervention, Comparison, and Outcome) to structure study inclusion and exclusion criteria (Table 1). Our population of interest was the general public or any subgroup, including health care professionals and students. We sought studies describing tools with gamification techniques or gamified elements, including gamified web-based quizzes to deliver informative or educative messages on vaccination. Posters, preprints, editorials, conference proceedings, news bulletins, and paper-based or board games were excluded. Our comparator was any control, including offering no education on vaccination or comparing participants before and after an intervention. Our outcomes of interest included common outcomes associated with vaccine uptake, namely knowledge (comprehension, understanding), attitudes (for or against vaccination), beliefs (perceived benefits, perceived risks), and behaviors toward vaccines (vaccination intention [ie, intention to get vaccinated or not get vaccinated] and vaccine uptake [ie, receiving or not receiving a vaccine]). We excluded papers that did not present the description or evaluation of a concrete digital gamified tool.

<table>
<thead>
<tr>
<th>Component</th>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
<th>Question related to the criteria</th>
</tr>
</thead>
</table>
| Type of report | - Original paper  
- Evaluated intervention or digital gamified tool | - Posters, preprints, and conference proceedings  
- Modeling or simulation study  
- Brochures  
- Editorials  
- Bulletins | Has the study or research described the development of the tool and evaluated it? |
| Population | - General public (any subgroup)  
- Professionals  
- Students | N/A | Who is the audience for whom the key message was intended? |
| Intervention | - Tools with gamification technique or gamified elements, including gamified web-based quizzes to deliver informative or educative messages on vaccination | - Any study or gamification tools not intended for vaccination/vaccine uptake  
- Studies or apps to reduce vaccine pain and fears and to report immunization status or record keeping, surveillance or vaccine coverage apps, contact-tracing or early detection apps  
- Paper games, board games (not digital)  
- Videos with no gamified element included | Does the study or tool aim to deliver an informative or educative message on vaccination? |
| Comparator | - Any control, including offering no education or no digital gamified tool | N/A | N/A |
| Outcome | - Common outcomes that encourage vaccine uptake: knowledge (comprehension, understanding), attitudes (for or against vaccination, beliefs [risk perception, etc], behaviors toward vaccines [vaccination intention [ie, intention to get vaccinated or not get vaccinated] and vaccine uptake [ie, receiving or not receiving a vaccine]) | - Outcomes not related to the encouragement of vaccine uptake | Has the study or tool been evaluated for the outcomes that encourage vaccine uptake? |

aN/A: not applicable.

For Google-searched digital gamified tools, our inclusion and exclusion criteria used the same specifications regarding
population and intervention. We did not apply comparison and outcome criteria to web-based tools because we did not expect these to report evaluation studies.

We reported this review according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (see the PRISMA checklist in Multimedia Appendix 2) [39]. We registered our protocol on the Open Science Framework [40].

**Expert Consultations**

After extracting information from peer-reviewed papers and tools identified via a Google search, we contacted experts in the field of digital gamified tools (eg, developers and researchers working on the topic in Canada and worldwide who were already known to the research team) to complement our online searches and ensure comprehensiveness. Specifically, we sent emails to 12 experts about the results of our searches and asked them to alert us to any games or papers we might have missed.

**Data Charting**

We developed a form in Microsoft Excel to guide the charting of data. We pretested and reviewed the form with team members to ensure we were accurately and adequately capturing relevant data. Data charting occurred independently with verification. Specifically, a reviewer (author HH) identified and screened all studies and digital gamified tools for their eligibility. Screening results were verified by a second reviewer (author DG). The data charting was then performed by a reviewer (HH) and again verified by a second reviewer (DG). Any conflicts throughout screening or data charting were resolved by a third reviewer (author ED). From the included papers, we charted data about (1) the type and design of study (developmental or evaluation study, user testing, randomized controlled trial, etc), (2) the vaccine(s) addressed (COVID-19, HPV, etc), (3) the purpose of the study or intervention, (4) the digital gamified tool platform (web based, native mobile app, mobile-enabled web app, virtual reality tool), and (5) the characteristics of study participants. For the evaluated interventions, we charted data about preselected outcomes that are widely used to predict health-related behaviors and to assess outcomes in studies of interventions about vaccination and immunization [11-14]. Specifically, we extracted data about the tools’ effects on knowledge, attitudes, beliefs (perceived benefits, perceived risks), and behavioral intentions. Emotional, cultural, and social factors can also influence a decision about vaccination [29,30]. Therefore, we also extracted data about other outcomes that the studies may have evaluated. Because we sought to understand all possible effects, we did not prespecify any of these as a primary outcome.

We organized the extracted data in tables and synthesized them descriptively.

**Quality Assessment**

To assess the quality of the studies that evaluated their interventions, we used the Mixed Methods Appraisal Tool (MMAT) developed by Pluye et al [41]. Two reviewers independently conducted the quality assessment, resolving disagreements through discussion until reaching a consensus. A third and a fourth reviewer (authors HH and HW) intervened to settle any remaining conflicts.

**Data Synthesis**

We summarized data using a narrative approach involving framework and content analysis. We classified each digital gamified tool platform using the 4 types of digital gamified tools: web-based tool, native mobile app, mobile-enabled web app, virtual reality tool. For the type of digital gamified tool, we classified web-based tools that explicitly noted their suitability for mobile use (eg, by smartphone or tablet) as mobile-enabled apps. We classified web-based tools without such an explicit statement as web based only, even though they may be functional on mobile devices. For the type and design of study, we grouped randomized designs together, including traditional randomized controlled trials with only 2 study arms and factorial designs with more than 2 study arms. Although these methods are not exactly the same, they all use randomization to minimize potential biases and are therefore functionally equivalent for our purposes of understanding what kinds of evaluations have been undertaken [42]. We summarized the main characteristics of tools, including PICO elements, in a tabular display. We used the PRISMA 2020 flowchart to describe the process of study selection [43].

**Results**

**Papers Identified and Scope of Literature**

We identified a total of 2082 records through database searches. After removing duplicates, we screened 1402 (67.3%) database records. Through Google searches, we identified 10 digital gamified tools and 2 papers. In a private browsing mode search, there was no change in search results. Of the 12 experts contacted, 2 (17%) responded and suggested 2 papers and 2 links, of which 1 (50%) digital gamified tool met the inclusion criteria and was included in our review. Through these methods, our final data set included 28 (2%) peer-reviewed papers and 11 digital gamified tools. Figure 1 shows our PRISMA diagram.
Of the 28 peer-reviewed papers, 7 (25%) explained the development of the tool, 16 (57%) described evaluation, and 2 (7%) reported both development and evaluation of the tool (Table 2). To report our results, we grouped studies together that reported the same tool, meaning 28 peer-reviewed papers reporting on 25 different tools. Of these 25 digital gamified tools, 11 (44%) were web-based tools, 7 (28%) mobile (native mobile or mobile-enabled web) apps, 6 (24%) virtual reality tools, and 1 (4%) offered in both mobile and web-based versions (for details, see Table 2). The most common single vaccines addressed in the tools were influenza (n=6, 24%, tools) and HPV (n=6, 24%, tools). Other tools addressed COVID-19 (n=2, 8%); measles, mumps, influenza, and smallpox (n=2, 8%); a hypothetical disease (n=2, 8%); other vaccine-preventable diseases (n=6, 24%); and the role vaccines play in preventing the spread of disease with no particular vaccine specified (n=1, 4%). Of the 10 digital gamified tools identified via a Google search and 1 suggested by the expert (a total of 11 digital gamified tools; see Table 3), the largest group (n=5, 45%) addressed COVID-19, and the rest were about other vaccine-preventable diseases. The 11 gamified elements identified in the Google search and expert feedback identified 6 types of gamified elements: reward points, serious games, physical trading cards, certificates, role-playing, and quizzes (see Table 3). The most common type was reward points, which appeared in 5 (45%) cases. Two cases used serious games, one case used physical trading cards and reward points, one case used certificates, one case used role-playing, and one case used quizzes. Additional characteristics of the studies included (eg, country of origin, sample size, participant characteristics) are detailed in Multimedia Appendix 3 [31,44-70]. The expanded versions of Table 2 [31,44-70] and Table 3 [71-81] are provided in Multimedia Appendix 4.
Table 2. General information about the studies.

<table>
<thead>
<tr>
<th>Type of study and author(s)</th>
<th>Type of digital gamified tool platform</th>
<th>Type of disease/vaccine</th>
<th>Type and design of study (development or evaluation, iterative design, randomized controlled trial, etc)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betsch and Böhm [44]</td>
<td>Web-based tool</td>
<td>Hypothetical</td>
<td>Evaluation: online experiment</td>
</tr>
<tr>
<td>Carolan et al [45]</td>
<td>Web-based tool</td>
<td>Measles, mumps, influenza, and smallpox</td>
<td>Evaluation: pre-post study</td>
</tr>
<tr>
<td>Cates et al [31]</td>
<td>Web-based tool</td>
<td>HPV&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Evaluation: pilot randomized controlled trial</td>
</tr>
<tr>
<td>Dale et al [46]</td>
<td>Native mobile app</td>
<td>Influenza</td>
<td>Evaluation: nonrandomized trial</td>
</tr>
<tr>
<td>Darville et al [47]</td>
<td>Web-based tool</td>
<td>HPV</td>
<td>Evaluation: randomized controlled trial</td>
</tr>
<tr>
<td>Eley et al [48], McNulty et al [49]</td>
<td>Web-based tool</td>
<td>Bacteria, vaccine-preventable disease</td>
<td>Evaluation: quantitative followed by qualitative research design</td>
</tr>
<tr>
<td>Fadda et al [50], Fadda et al [51]</td>
<td>Native mobile app</td>
<td>MMR vaccines</td>
<td>Evaluation: mixed methods research design</td>
</tr>
<tr>
<td>Ibuka et al [52]</td>
<td>Web-based tool</td>
<td>Hypothetical disease</td>
<td>Evaluation: experimental design</td>
</tr>
<tr>
<td>Lee et al [54]</td>
<td>Native mobile app</td>
<td>Influenza</td>
<td>Evaluation: randomized controlled trial</td>
</tr>
<tr>
<td>Mitchell et al [55], Laplana [56]</td>
<td>Web-based tool</td>
<td>Influenza</td>
<td>Evaluation: pre-post study</td>
</tr>
<tr>
<td>Mottelson et al [57]</td>
<td>Virtual reality tool</td>
<td>COVID-19</td>
<td>Evaluation: randomized controlled trial (2x2 factorial design)</td>
</tr>
<tr>
<td>Nowak et al [58]</td>
<td>Virtual reality tool</td>
<td>Influenza</td>
<td>Evaluation: one-way between-subjects design with random assignment</td>
</tr>
<tr>
<td>Real et al [59]</td>
<td>Virtual reality tool</td>
<td>Influenza</td>
<td>Evaluation: quasi-randomized controlled trial&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Woodall et al [60]</td>
<td>Mobile-enabled web app</td>
<td>HPV</td>
<td>Evaluation: clinic-cluster randomized trial</td>
</tr>
<tr>
<td>Wandeweerdt et al [61]</td>
<td>Virtual reality tool</td>
<td>COVID-19</td>
<td>Evaluation: randomized controlled trial</td>
</tr>
<tr>
<td><strong>Development studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amresh et al [62]</td>
<td>Web-based tool</td>
<td>HPV</td>
<td>Development: iterative design</td>
</tr>
<tr>
<td>Bertozzi et al [63] (data extracted for the game related to vaccines)</td>
<td>Web-based tool</td>
<td>Influenza</td>
<td>Development: iterative design</td>
</tr>
<tr>
<td>Carolan et al [64]</td>
<td>Web-based tool</td>
<td>Measles, mumps, influenza, and smallpox</td>
<td>Development: iterative design</td>
</tr>
<tr>
<td>de Araujo Lima et al [66]</td>
<td>Native mobile app</td>
<td>Vaccine-preventable diseases</td>
<td>Development: heuristic evaluation by users, content evaluation by experts</td>
</tr>
<tr>
<td>Kafai et al [65]</td>
<td>Virtual reality</td>
<td>Dragon swooping cough virus to reflect real-life features of infectious viruses, such as Ebola.</td>
<td>Development: user feedback via surveys (asking users questions) and log files (observing user behaviors)</td>
</tr>
<tr>
<td>Real et al [67]</td>
<td>Native mobile app</td>
<td>HPV</td>
<td>Development: usability testing</td>
</tr>
<tr>
<td>Streuli et al [68]</td>
<td>Virtual reality</td>
<td>Pediatric vaccines</td>
<td>Development: Community-based participatory research and co-design</td>
</tr>
<tr>
<td><strong>Development and evaluation studies</strong></td>
<td>Mobile or web app (multiple formats available)</td>
<td>Hepatitis B</td>
<td>Development and evaluation: Participatory Action Research</td>
</tr>
<tr>
<td>Ruiz-López et al [70]</td>
<td>Native mobile app</td>
<td>HPV</td>
<td>Development and evaluation: Iterative design and evaluation via questionnaire</td>
</tr>
</tbody>
</table>

<sup>a</sup>HPV: human papillomavirus.
Allocation to a study arm was performed according to work schedules, which are often arbitrary. We therefore considered this quasi-randomization.

Table 3. Tools from Google search and expert suggestions.

<table>
<thead>
<tr>
<th>Digital gamified tool</th>
<th>Type of disease/vaccine</th>
<th>Type of digital gamified tool platform</th>
<th>Gamification elements (eg, rewards, role-playing, leaderboard, serious game)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antidote COVID-19 [71]</td>
<td>COVID-19</td>
<td>Native mobile app</td>
<td>Reward points</td>
</tr>
<tr>
<td>The Vaccination Game [72]</td>
<td>H1N1 and influenza</td>
<td>Web-based tool</td>
<td>Serious game</td>
</tr>
<tr>
<td>Help take down COVID-zilla! [73]</td>
<td>COVID-19</td>
<td>Web-based tool</td>
<td>Role-playing</td>
</tr>
<tr>
<td>Just the Vax! [74]</td>
<td>Vaccine-preventable disease</td>
<td>Web-based tool</td>
<td>Reward points</td>
</tr>
<tr>
<td>COVID Invaders [75]</td>
<td>COVID-19</td>
<td>Web-based tool</td>
<td>Reward points</td>
</tr>
<tr>
<td>Vax Pack Hero [76]</td>
<td>Vaccine-preventable disease</td>
<td>Web-based tool</td>
<td>Reward points and physical trading cards</td>
</tr>
<tr>
<td>Flu's Clues [77]</td>
<td>Influenza</td>
<td>Web-based tool</td>
<td>Certificate of completion for solving the influenza mystery</td>
</tr>
<tr>
<td>Virus Fighter [78]</td>
<td>COVID-19, influenza, Ebola, measles</td>
<td>Web-based tool</td>
<td>Serious game</td>
</tr>
<tr>
<td>Immunization411: for pre-teens and teens’ online training [79]</td>
<td>Tdap meningococcal vaccine, varicella, HPV(^a), influenza</td>
<td>Web-based tool</td>
<td>Reward points</td>
</tr>
<tr>
<td>COVID Chronicles [80]</td>
<td>COVID-19</td>
<td>Web-based tool</td>
<td>Reward points</td>
</tr>
<tr>
<td>I Boost(^b) [81]</td>
<td>Vaccine-preventable disease</td>
<td>Web-based tool</td>
<td>Quiz</td>
</tr>
</tbody>
</table>

\(^a\)HPV: human papillomavirus.

\(^b\)Suggested by an expert.

The studies were conducted in 26 different countries, with the majority of studies coming from the United States (n=13, 46%, studies) and the United Kingdom (n=5, 18%, studies). Study populations included students at various levels (elementary school to college, specialty programs, eg, nursing and pediatric residency), parents of vaccine-eligible children, adults from the general population, members of particular sociocultural communities (eg, immigrants, Indigenous peoples), and convenience samples, such as players of a game, attendees of a conference, and employees of an organization. Sample sizes ranged from 8 to 50,286. Whenever papers reported study participant characteristics such as age, sex, gender, ethnocultural identity, or socioeconomic levels, we extracted summary data, as shown in Multimedia Appendix 3.

**Reported Effects of Evaluated Interventions**

In total, 18 (64%) of 28 studies evaluated at least 1 of our outcomes of interest, while 11 (39%) studies reported the effects of the evaluated interventions on more than 1 outcome of interest. Summarized outcomes and their MMAT quality assessments are shown in Table 4. Multimedia Appendix 5 provides full details.
<table>
<thead>
<tr>
<th>Type of digital gamified tool platform and study</th>
<th>Knowledge (comprehension/understanding, etc)</th>
<th>Attitudes (for/against vaccination, etc)</th>
<th>Beliefs (risk perceptions, etc)</th>
<th>Behavioral intentions (getting vaccinated or not, etc)</th>
<th>Others (eg, emotions)</th>
<th>MMAT\textsuperscript{a} quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web-based tool</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betsch and Böhm [44]</td>
<td>— (b)</td>
<td>Negative vaccine attitudes with compulsory vaccination</td>
<td>—</td>
<td>Decreased vaccine uptake with compulsory vaccination</td>
<td>Increased level of anger with compulsory vaccination</td>
<td>60% quality criteria met</td>
</tr>
<tr>
<td>Carolan et al [45]</td>
<td>—</td>
<td>No significant effect on attitudes towards vaccination</td>
<td>—</td>
<td>—</td>
<td>Increased confidence in information needs</td>
<td>80% quality criteria met</td>
</tr>
<tr>
<td>Cates et al [31]</td>
<td>Increase in knowledge about immunization</td>
<td>—</td>
<td>—</td>
<td>Positive increase in intentions to vaccinate</td>
<td>Increase in vaccination self-efficacy, decisional balance towards vaccination</td>
<td>100% quality criteria met</td>
</tr>
<tr>
<td>Darville et al [47]</td>
<td>—</td>
<td>—</td>
<td>Positive effects on beliefs towards vaccination</td>
<td>Increase in intentions to vaccinate</td>
<td>—</td>
<td>60% quality criteria met</td>
</tr>
<tr>
<td>Eley et al [48], McNulty et al [49]</td>
<td>Improvements in knowledge about immunization</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>100% quality criteria met</td>
</tr>
<tr>
<td>Ibuka et al [52]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Free riding in vaccination decisions decreases vaccine acceptance</td>
<td>—</td>
<td>80% quality criteria met</td>
</tr>
<tr>
<td>Kaufman and Flanagan [53]</td>
<td>The digital version of the game was less effective at facilitating learning</td>
<td>The digital version of the game was less effective at attitude change</td>
<td>—</td>
<td>—</td>
<td>The digital version of the game was perceived to be complicated to use</td>
<td>20% quality criteria met</td>
</tr>
<tr>
<td>Mitchell et al [55], Laplana [56]</td>
<td>Increase in knowledge</td>
<td>Positive increase in attitudes for vaccination</td>
<td>—</td>
<td>Increase in vaccine uptake after accessing the game</td>
<td>—</td>
<td>80% quality criteria met (Mitchell et al [55])</td>
</tr>
<tr>
<td><strong>Mobile app</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dale et al [46]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Positive increase in intentions to vaccinate</td>
<td>—</td>
<td>80% quality criteria met</td>
</tr>
<tr>
<td>Fadda et al [50], Fadda et al [51]</td>
<td>Improvements in knowledge about immunization</td>
<td>—</td>
<td>—</td>
<td>Increase in intentions to vaccinate</td>
<td>Increase in psychological empowerment and confidence in the decision</td>
<td>80% quality criteria met (Fadda et al [50], Fadda et al [51])</td>
</tr>
<tr>
<td>Lee et al [54]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Increase in intentions to vaccinate</td>
<td>—</td>
<td>80% quality criteria met</td>
</tr>
<tr>
<td>Woodall et al [60]</td>
<td>—</td>
<td>—</td>
<td>Increase in beliefs towards vaccination</td>
<td>Increase in intentions to vaccinate</td>
<td>—</td>
<td>40% quality criteria met</td>
</tr>
<tr>
<td>Ruiz-López et al [70]</td>
<td>Increase in knowledge after playing the game</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>100% quality criteria met</td>
</tr>
<tr>
<td><strong>Virtual reality tool</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\text{MMAT}^a\) represents the quality score for each study.
**Effects on Knowledge (Includes Comprehension/Understanding, etc)**

Overall, the 28 included studies suggested that digital gamified tools may positively influence knowledge. Of 7 (25%) studies that assessed knowledge, 6 (86%) showed an increase in knowledge about immunization in general [31,48,51,55,69,70]. All these 6 (86%) studies were of high quality (≥80%). One study of low quality (≤25%) reported that a digital game is less effective at increasing knowledge compared to its original board game format [53]. When considering only the high-quality (≥80%) studies, we observed that digital gamified tools are associated with increased knowledge.

**Effects on Attitudes (for or Against Vaccination)**

Overall, digital gamified tools appeared to have mixed effects on attitudes toward vaccination. Of 5 (18%) of 28 studies that assessed attitudes, 2 (40%), one of high quality (≥80%) and the other of medium quality (60%), showed an increase in positive attitudes toward vaccination [55,59]. In addition, 2 (40%) studies, one of high quality (≥80%) and the other of low quality (20%), reported no or less effect on attitudes toward vaccination [45,53], and 1 (20%) study comparing voluntary and compulsory vaccines in a game context showed negative attitudes regarding compulsory vaccination [44]. When considering only the high-quality (≥80%) studies, we observed inconsistent effects on attitudes.

**Effects on Beliefs (Perceived Benefits, Perceived Risks)**

Overall, digital gamified tools demonstrated positive effects on beliefs toward vaccination. In total, 3 (11%) of 28 studies, 1 (33%) of high quality (100%) and 2 (67%) of medium quality (60% and 40%), evaluated the effects of digital gamified tools on beliefs toward vaccination. All 3 (100%) studies showed positive effects on beliefs toward vaccination [47,58,60]. When considering only the high-quality (≥80%) studies, we observed that digital gamified tools are associated with positive beliefs about vaccines.

**Effects on Behavioral Intentions**

Overall, the 28 included studies suggested that digital gamified tools may positively influence intentions to receive vaccines. In total, 11 (39%) studies evaluated the effects of digital gamified tools on behavioral intentions with regard to vaccines. Of these 11 studies, 1 (9%) of medium quality (60%) showed a decrease in vaccination intention when compulsory vaccination was introduced within a game context [44], whereas 10 (91%) studies, 3 (30%) of medium quality (60% and 40%) and 7 (70%) of high quality (≥75%), showed increased intentions to vaccinate [31,46,47,51,54,55,57,58,60,61]. When considering only the high-quality (≥80%) studies, digital gamified tools appeared to be consistently associated with increased vaccination intention.

**Other Outcomes**

In total, 9 (32%) of 28 studies have also evaluated the effects of digital gamified tools on other outcomes. Of these, 4 (44%) studies reported an increase in confidence in vaccines (medium quality=40%) [60], confidence in information needs (high quality=80%) [45], decisional balance in support of vaccination (high quality=100%) [31], and confidence in vaccine decisions (high quality=80%) [50]. In addition, 1 (11%) study of high
quality (80%) reported an increase in empathy toward those vulnerable to COVID-19 and vaccination recommendations [57], and 2 (22%) studies of high quality (100% and 80%) reported an increase in vaccination self-efficacy and readiness [31,57]. An increase in psychological empowerment (high quality=80%) [51] and in emotions such as anger toward compulsory vaccination (medium quality=60%) [44] was also reported by 2 (22%) studies. One study of high quality (80%) reported that the concept of free riding decreases vaccine acceptance [52], whereas another study of high quality (100%) reported that virtual reality intervention increases collective responsibility [61]. When considering only the high-quality (>80%) studies, we observed a variety of positive effects associated with digital gamified tools, including confidence in vaccines, confidence in decisions about vaccines, empathy toward vulnerable people, collective responsibility, psychological empowerment, and vaccination self-efficacy and readiness.

**Effects of the Platform (Web Based, Mobile, Virtual Reality)**

The study designs of the 28 included papers did not permit us to formally compare the effects of different platforms in a robust way. Upon inspection, there did not appear to be a strong effect of the platform. In other words, we did not observe evidence in favor of web-based, mobile, or virtual reality apps over the other 2 types of apps.

**Discussion**

**Principal Findings**

The broad objective of this scoping review was to map the state of the science regarding digital gamified tools and their effects. In other words, we wished to answer a common question at the intersection of public health and digital health: does gamification encourage vaccination and influence knowledge, attitudes, beliefs, and behaviors related to vaccination? By mapping both published literature and tools currently available online, we observed 2 principal findings.

First, our results suggest that gamification can increase predictors of vaccine uptake, such as knowledge, attitudes, beliefs, behaviors, and vaccination intention. This finding is similar to the findings of a previous review by Montagni et al [82] suggesting that gamification can contribute to changed behaviors and improved knowledge of vaccination. Similarly, other reviews have suggested the potential benefits of gamification for non–vaccination-related behavior change, such as a systematic review suggesting that gamification interventions could be a feasible way to improve health-related outcomes among cancer survivors [83] and another review suggesting their effectiveness in improving physical activity [84]. Such previous work became even more relevant during the COVID-19 pandemic, as many jurisdictions sought to optimize vaccine uptake in the context of an “infodemic” (ie, overabundance of information, true, false, and misleading, about the pandemic and recommended preventive behaviors) [85]. Half of the digital gamified tools identified in our web search addressed COVID-19, suggesting an active interest in using a gamified approach in the pandemic context. Recent research by Plechatá et al [86] published after our data extraction steps were complete suggested that explaining the concept of herd immunity with gamification has a positive impact on the COVID-19 vaccination intention.

Second, our review suggests that although gamification has the potential to enhance the impact of education strategies, gamified tools alone may not wholly address gaps in vaccine acceptance and uptake. Although some of the identified tools did increase vaccination, the increases did not fully close gaps between previous and desired vaccine uptake. This finding aligns with those of Tozzi et al [87], which suggested that promising results could be achieved by combining gamification with educative and informative tools to improve immunization programs. This finding also aligns with previous reviews suggesting the use of digital gamified interventions as a public health tool of interest in enhancing vaccine uptake [82,88]. Further research published by Real et al [89] after our systematic search similarly observed that integrating gamification, such as virtual reality, in training modules enhances uptake of the HPV vaccine. Integrating gamified features may work because they make digital tools acceptable and more fun to use and may reduce the chances of people feeling pushed toward vaccination. In parallel, gamification may be a promising strategy for increasing knowledge, skills, and confidence among health professionals engaging in discussions about vaccines with their patients [90,91].

In addition to these findings drawn directly from our review of the included tools, we offer a broader observation based on the contents of this scoping review, along with the larger landscape of vaccine acceptance research: context is key. Although an engaging approach may work for some groups or in some situations, it may be less well accepted among other groups and in other situations. For instance, a casual and approachable style of communication will work for the younger audience to convey vaccine information but might be deemed insufficient to health care professionals in a more formal setting, such as hospitals. A good understanding of the factors associated with low vaccine acceptance at the local level is needed prior to developing gamified tools [92]. Future research in this area should consider possible contextual factors, such as local culture, social and demographic characteristics of users, and different influences on vaccine hesitancy and acceptance in different regions. To help better match games to the context(s) in which they will be played, when developing games, developers and researchers may wish to consider involving potential players from different contexts early and often. This aligns with previous work [93,94] suggesting that involving users earlier in developing tools may help in designing interventions suitable for a targeted context. One of the examples in our review was an intervention by Cates et al [31] designed to explain HPV vaccines to teenagers using a “secret garden” theme. Involving potential game players early in the development of the game may have contributed toward its positive effects on vaccination intention.

The implications of this research extend beyond the immediate reported effects of gamified tools and delve into the strategic dimensions of public health policy and communication efforts. Considering the insights gleaned from the findings, this study
supports a comprehensive and well-informed approach to integrating gamification into strategies for promoting vaccination. As gamification continues to demonstrate its potential in enhancing vaccine uptake, it is crucial to navigate this terrain thoughtfully, considering the various factors that influence its impact. This includes not only the technological and behavioral aspects but also the larger sociocultural context in which vaccination decisions are made. Therefore, our study emphasizes the importance of a comprehensive approach that fosters a mutually beneficial relationship between technological innovation, evidence-based strategies, and an intricate understanding of local contexts. This approach has the potential to make gamification a sustainable and adaptable tool in the arsenal of public health interventions, rather than just a passing trend.

The review does not find a clear advantage for any platform in terms of reported effects. It was challenging to measure the impact of the platforms on behavioral outcomes and calls for more focused research to better understand the specific elements within each platform that drive behavior change. In essence, our study suggests that the reported effects of an app may not be solely determined by its platform but rather by the strategic incorporation of mechanics and elements that facilitate the desired behavior change.

Gamification can influence knowledge, attitudes, and beliefs about vaccines, which can affect vaccine uptake. This is consistent with theories of change proposing that cognitive changes can lead to behavioral outcomes. Although our study mainly examines the immediate effects of gamification on these cognitive aspects, it also offers some implications for using gamification as a potentially viable strategy to improve vaccine acceptance.

Strengths and Limitations

Our study has 5 main limitations. First, because we aimed to capture all relevant evidence and examples, as is typical in a scoping review, we included a broad range of study designs and did not draw conclusions about the relative advantages or disadvantages of different game platforms and features. Given the rapid growth within this field of research, it would be difficult to truly prioritize evidence according to quality criteria at this point. In the future, it may be possible to conduct a systematic review and meta-analysis, restricting included studies to randomized experiments or randomized controlled trials. Such future work may include approaches such as a network meta-analysis to allow for comparison of the effects of different game types or game features. Based on the existing literature, it is difficult to conclude whether certain games are more or less likely to achieve their aims. Second, our results may be influenced by publication bias. It is possible that groups that have developed digital gamified tools that showed disappointing results simply did not publish their studies. This bias could lead to an overestimation of the reported effects of these tools. This highlights the importance of further research to fully understand the real impact of these tools and thus accurately inform policy decisions about the development and use of these tools. Third, and related to the previous 2 points, the rapid growth in this area may mean that we missed more recent evidence in literature published after January 2022 and web searches after May 2022. Fourth, the majority of digital gamified tools on vaccination represented in publications and online were developed in high-income countries. This finding aligns with the findings of previous work by Ohannessian et al [88], who also reported a predominance of high-income countries. This may reflect more widespread internet access and resources for developing digital gamified tools in high-income countries. It may also reflect publication bias in the scientific literature (ie, there may be fewer papers written about digital gamified tools in lower-income countries) and online (ie, tools developed and published in lower-income countries may not be ranked highly by search engines and therefore may not have appeared in our web searches). Tools developed in lower-income countries may also take different forms; for example, they may be text message–based interventions (with or without gamification) rather than web-based tools and therefore would be less likely to be identified in web searches. Analog games from high-income countries were similarly excluded from the scope of our study [95]. Nondigital games, such as board and card games, have demonstrated positive impacts on educational knowledge, cognitive function, and social interactions [96,97]. Such games can support diverse learning across subjects and settings, fostering interactions that develop skills, such as computational thinking and teamwork, and have positive impacts on academic achievement and vocabulary acquisition compared to digital games [97-99]. We restricted our scoping review to digital gamified tools because the review was intended to provide an evidence base for digital game development. Although nondigital games are also potentially useful interventions, the implementation and distribution of such interventions is more challenging, especially in a geographically dispersed country, such as Canada. Fifth, and finally, as we used Google and private browsing in Google, there may be a possibility that different search engines would provide different results.

This study also has 2 main strengths. First, by systematically examining the current literature and currently available tools online, we were able to offer an updated overview of the potential effects of including gamification in digital tools about vaccination. Second, by conducting a scoping review to broadly map the literature, future work can more easily identify and select key outcomes for systematic reviews and meta-analyses in this domain.

Conclusion

Digital gamified tools have the potential to improve vaccine uptake by increasing knowledge and promoting positive attitudes, beliefs, behaviors, and vaccination intention. Further evaluations of these innovative digital tools, including head-to-head comparisons of different features and different platforms, will add more knowledge about what works and what does not in order to achieve public health goals more efficiently. In the wider context of health policy, digital gamified tools may be useful components of multifaceted strategies to improve vaccination rates throughout society.
Acknowledgments
This work was supported by a research grant from the Canadian Institutes of Health Research (GA3177725). The authors gratefully acknowledge the assistance of Frédéric Bergeron (librarian) for reviewing the search strategy. We appreciate the assistance of Charles Racine and Crésence Joëlle Mefou Tasong in conducting the quality appraisal of the studies that evaluated their interventions.

Authors' Contributions
All authors provided substantial contributions to this paper’s conception and edits and approved the final version of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Search strategy.  
[DOCX File, 20 KB - games_v12i1e47257_app1.docx ]

Multimedia Appendix 2
PRISMA-ScR checklist.  
[PDF File (Adobe PDF File), 498 KB - games_v12i1e47257_app2.pdf ]

Multimedia Appendix 3
Characteristics of the studies included in the review.  
[DOCX File, 25 KB - games_v12i1e47257_app3.docx ]

Multimedia Appendix 4
Expanded version of Table 2 (general information about the studies) and Table 3 (tools from Google search and expert suggestions).  
[DOCX File, 28 KB - games_v12i1e47257_app4.docx ]

Multimedia Appendix 5
[XLSX File (Microsoft Excel File), 92 KB - games_v12i1e47257_app5.xlsx ]

References
4. Ten health issues WHO will tackle this year. World Health Organization. URL: http://tinyurl.com/39cd7uf [accessed 2021-10-04]


39. PRISMA for scoping reviews. PRISMA. URL: http://tinyurl.com/23j8b9aw [accessed 2024-02-10]
40. Gamification tool–environment scan. OSF Registries. URL: https://osf.io/nv8af [accessed 2024-02-09]


62. Amresh A, Chia-Chen A, Baron C. A game based intervention to promote HPV vaccination among adolescents. 2019 Presented at: SeGAH 2019: IEEE 7th International Conference on Serious Games and Applications for Health; August 5-7, 2019; Kyoto, Japan. [doi: 10.1109/segh.2019.8882459]


72. The vaccination game. mrcwimm. URL: https://mrcwimm.itch.io/the-vaccination-game [accessed 2024-02-09]

73. Help take down COVID-zilla!: protect your friends and family. Alberta Health Services. URL: http://tinyurl.com/yp82u75j [accessed 2024-02-09]

74. Just the vax!. Vaccine Education Center. URL: https://media.chop.edu/data/files/vaccine-trivia-game/index.html [accessed 2024-02-09]

75. COVID invaders. Get One Desk. URL: https://www.getonedesk.com/covid-invaders [accessed 2024-02-09]


78. Virus fighter. Virus Fighter. URL: https://www.virusfighter.org/ [accessed 2024-02-09]


80. COVID chronicles. COVID Chronicles. URL: http://tinyurl.com/vc73frhr [accessed 2024-02-09]

82. Montagni I, Mabchour I, Tzourio C. Digital gamification to enhance vaccine knowledge and uptake: scoping review. JMIR Serious Games 2020 May 18;8(2):e16983 [FREE Full text] [doi: 10.2196/16983] [Medline: 32348271]


Abbreviations

HPV: human papillomavirus
MMAT: Mixed Methods Appraisal Tool
PICO: Population, Intervention, Comparison, and Outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
©Hina Hakim, S Michelle Driedger, Dominique Gagnon, Julien Chevrier, Geneviève Roch, Eve Dubé, Holly O Witteman. Originally published in JMIR Serious Games (https://games.jmir.org), 29.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.
Health Care Professional–Supported Co-Design of a Mime Therapy–Based Serious Game for Facial Rehabilitation

Daniel Lima Sousa1, MSc; Silmar Teixeira1, PhD; José Everton Fontenele1, MSc; Renato Mendes Santos1, MSc; Leynilson Pereira1, MSc; Rodrigo Baluz1,2, PhD; Victor Hugo Bastos1, PhD; Ariel Soares Teles1,3, PhD

1Parnaíba Delta Federal University, Parnaíba, Brazil
2State University of Piauí, Campus Parnaíba, Parnaíba, Brazil
3Federal Institute of Maranhão, Campus Araioses, Araioses, Brazil

Corresponding Author:
Ariel Soares Teles, PhD
Federal Institute of Maranhão, Campus Araioses
Rua José de Alencar, S/N
Bairro Cumprida
Araioses, 65570-000
Brazil
Phone: 55 86995501313
Email: ariel.teles@ifma.edu.br

Abstract
This research letter presents the co-design process for RG4Face, a mime therapy–based serious game that uses computer vision for human facial movement recognition and estimation to help health care professionals and patients in the facial rehabilitation process.

(JMIR Serious Games 2024;12:e52661) doi:10.2196/52661

KEYWORDS
serious game; serious games; facial recognition; face estimation; computer vision; facial rehabilitation; face; rehabilitation; physiotherapy; mime therapy; co-design; human face estimation; facial palsy; facial paralysis; motor rehabilitation; exergame; physiotherapists; psychologists; participation

Methods

Ethical Considerations
This study was approved by the Research Ethics Committee of Universidade Federal do Delta do Parnaíba (5.632.311). The first author (DLS) provided explicit consent for use of his image in Multimedia Appendices 1 and 2.

Study Design
To develop RG4Face, a co-design procedure (Figure 1) was conducted with physiotherapists (n=16) and psychologists (n=5; Multimedia Appendix 3) to obtain the necessary knowledge on the game requirements.

In the first stage, a version of the game was developed with an initial idea (Multimedia Appendix 1). In the second, we recruited physiotherapists and psychologists to participate in co-design meetings (August to November 2022) and answer a questionnaire. We then presented the game to the participants and allowed them to make suggestions. The prototype was
essential to encouraging participation during meetings. In total, 5 meetings were held—4 with physiotherapists and 1 with psychologists. The main activities of the meetings were brainstorming sessions, in which the generation of game requirements was encouraged for their incorporation into visual elements, gamification, and game mechanics. Meeting results allowed for the creation of a list of requirements. As a third stage, we are concluding the implementation of RG4Face based on the produced requirements. The game code was implemented in JavaScript to provide new features for facial rehabilitation via the Rehabilite Game platform [6].

Figure 1. Co-design timeline.

Results

Per its initial conception, RG4Face uses computer vision (via a camera) for capturing, recognizing, and estimating human facial movements. The game prototype was implemented via the MediaPipe face mesh [7] to enable the recognition and use of 1 movement (eg, raising eyebrows; ie, frontal muscle) to control game elements. The game involves a spaceship moving horizontally across the bottom of the captured video window and firing a projectile when face movement is detected. The main objective is to hit triangles that randomly appear on the player's face.

Table 1 presents participants’ suggestions during co-design, game requirements, and rationales.

Table 1. Participants’ suggestions during co-design, game requirements, and rationales.

RG4Face is in the testing phase and, prior to evaluations, can recognize 6 movements used in mime therapy to improve facial muscle strength and mobility (Multimedia Appendix 2). To implement the recognition of these movements, MediaPipe was used [7]. The face mesh model allows for the real-time tracking of 468 3D landmarks on the human face that represent important facial features (eg, eyes, eyebrows, nose, and mouth). Distances between landmarks are calculated to recognize movements.

RG4Face provides a mirror therapy feature [8], which can mirror the healthy side of the face to create a visual illusion that can help reduce pain and improve function. RG4Face allows for parameter adjustment on the Rehabilite Game platform. Health care professionals can choose specific game mechanics for each rehabilitation case, thereby customizing the game according to patients’ needs and difficulties.
Table 1. Functional and nonfunctional game requirements from the co-design procedure.

<table>
<thead>
<tr>
<th>Participants’ suggestions</th>
<th>Refined requirement</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Improve the game scenario</td>
<td>• Improve game colors and elements: border, ship, projectiles, and collision</td>
<td>• Enable the game to become more attractive and stimulating</td>
</tr>
<tr>
<td>• Improve the representation of the ship and projectiles</td>
<td>• Provide difficulty levels</td>
<td>• Gamification for each level, depending on the patient's condition</td>
</tr>
<tr>
<td>• Choose attractive colors and contrasts</td>
<td>• Implement a mirror therapy simulation</td>
<td>• Patients with Bell palsy can benefit from it</td>
</tr>
<tr>
<td>• Include levels with difficulty levels</td>
<td>• Create a scoring and bonus system</td>
<td>• Increase patients’ adherence to and engagement with treatment</td>
</tr>
<tr>
<td>• Provide an option of mirror therapy for the game</td>
<td>• Implementation of sensitivity levels for motion recognition</td>
<td>• The level of sensitivity respects the movement capacity of each patient</td>
</tr>
<tr>
<td>• Implement better game mechanics for rewards</td>
<td>• Provide in-game metrics</td>
<td>• They are interesting for the health care professional to follow the patient's progress</td>
</tr>
<tr>
<td>• Promote progression in the game</td>
<td>• Allow game sound to be optional (ie, turn off the sound)</td>
<td>• The sound may be unnecessary for some patients</td>
</tr>
<tr>
<td>• Movement sensitivity must be customized according to the patient's degree of disability</td>
<td>• Make game screen full, automatically adjusting to the aspect ratio</td>
<td>• Game elements should be clearly visible</td>
</tr>
<tr>
<td>• Create metrics on the game platform to monitor the rehabilitation process</td>
<td>• Perform a prior calibration of the player’s face</td>
<td>• Adjustment of the distance between player’s face and screen, in addition to improving movement recognition</td>
</tr>
<tr>
<td>• To avoid causing botheration to some types of patients, the sound should be optional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Consider visual acuity of the players</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The game scenario should be full screen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Head movement should not influence the game</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Calibration is essential to avoid false positives and false negatives of movements</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

We co-designed a serious game for facial rehabilitation that represents a potential new approach to improving patients’ adherence to facial rehabilitation. The co-design procedure allowed stakeholders to participate in defining game requirements, thereby empowering the tool to meet the needs and expectations of patients and be more engaging and motivating.

Although there are studies that focus on games for rehabilitating specific parts of the face (eg, eyes [9] and mouth [10]), to our knowledge, no serious game for facial rehabilitation has been proposed that can recognize the face movements used in mime therapy. This study proposes the first such exergame.

Our results demonstrate that the co-design approach was effective for creating a serious game with the potential to meet patients’ needs. We plan to evaluate the game with health care professionals, healthy participants, and patients with facial paralysis.

Acknowledgments

This work was supported by the Federal Institute of Maranhão, Coordination for the Improvement of Higher Education Personnel (CAPES; finance code 001), National Council for Scientific and Technological Development (CNPq; grants 308736/2022-2 and 308059/2022-0), and State Funding Agency of Maranhão (FAPEMA; grant UNIVERSAL-06123/22).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Video presentation with the game prototype before the co-design procedure.

[MP4 File (MP4 Video), 7354 KB - games_v12i1e52661_app1.mp4]

Multimedia Appendix 2

https://games.jmir.org/2024/1/e52661
Video presentation with the game after implementing requirements from the co-design procedure.

[MP4 File (MP4 Video), 55743 KB - games_v12i1e52661_app2.mp4 ]

Multimedia Appendix 3

Demographic characteristics of participants.

[DOCX File , 15 KB - games_v12i1e52661_app3.docx ]

References
