Exploring Usability in Virtual Reality: A Systematic Review of Evaluation Methods to Enhance Learning in Immersive Environments

Irvin Yair Acuña Mendoza[1], Collins Pool Vieira Abad[2], Mejía Cabrera Heber Ivan[3]

[1] School of Systems Engineering, Señor de Sipán University, Chiclayo 14000, Peru

*** Correspondence:**
amendozairvinya@uss.edu.pe
vabadcollinspoo@uss.edu.pe
hmejiac@uss.edu.pe

Keywords: System Usability, Virtual Reality, Learning.

Abstract

This systematic review explores the methods used to evaluate the usability of virtual reality (VR) applications designed for immersive learning environments. The main objective is to analyze scientific studies to identify the most effective techniques and metrics for assessing usability in educational VR contexts. Following PRISMA guidelines, a search was conducted in the Scopus and Web of Science databases for publications from 2020 to 2025. Out of an initial total of 393 records, 30 studies were selected for in-depth analysis, applying specific inclusion and exclusion criteria.The results indicate a strong preference for mixed-methods approaches (used in 80% of the studies), combining quantitative instruments such as the System Usability Scale (SUS) and NASA-TLX, with qualitative techniques such as interviews and observation. The SUS was the most commonly used standardized tool, appearing in 47% of the reviewed articles. Key objective metrics identified include: task completion rate (used in 20 studies), task time (42% of the studies), and error rate (11 studies). Although a generally positive correlation was observed between users' subjective perception of usability and their objective performance, factors such as VR-induced motion sickness (cybersickness) and cognitive load introduced certain inconsistencies. The study concludes that while there is a general consensus on the use of mixed methods, there remains a significant gap in the adoption of robust international standards such as ISO 9241.

1 Introduction

Achieving educational quality in modern society depends largely on the methods of learning. In this regard, Smart Learning Environments (SLEs) have emerged, designed to create effective learning settings gradually and sustainably through the application of technology (Maulidiya et al., 2024). At the same time, immersive environments have evolved significantly, encompassing a wide range of contexts such as traditional classrooms, online platforms, and virtual reality (VR) environments, fostering more versatile teaching methods (Papaioannou et al., 2023). Within this evolution, VR has emerged as an innovative tool for learning, offering immersive and interactive experiences that engage students.

Thanks to its ability to simulate three-dimensional scenarios analogous to real-life situations, VR facilitates a deeper and more practical understanding of complex concepts, especially in fields where traditional teaching methods fall short (Sulisworo et al., 2023). Recent studies highlight that VR increases motivation and academic performance by actively involving students in the learning process (Makransky and Lilleholt, 2018). However, the effective use of virtual reality in educational contexts depends on the usability of the VR application how intuitive, accessible, and efficient the experience is for students.

This is why the lack of standardized methods to evaluate the usability of VR applications represents a critical barrier to their adoption and effective development in learning environments. One of the main issues identified is the limited number of validated evaluation models that assess usability in VR applications for education (Sutcliffe and Gault, 2004). For instance, studies have shown that poorly designed virtual environments can lead to negative effects such as cybersickness, visual fatigue, and user frustration, which in turn limit their potential as learning tools (Kennedy et al., 1993).

This, in turn, increases the need for measurement instruments that incorporate usability criteria in VR applications (Johnson, 2005). Furthermore, the implementation of VR technologies in education faces challenges related to the availability of technological resources. The lack of adequate infrastructure and training in these technologies limits their integration into educational settings.

Evaluating the usability of educational tools is crucial to improving the quality of applications and their impact on the teaching-learning process. Although usability analysis tools exist, they are not fully focused on assessing VR environments within educational contexts. Therefore, there is a growing need to develop specialized methods for evaluating usability in VR (Paz et al., 2015). The standardization of usability evaluation methods is essential to ensure a better learning experience.

The lack of systematic research analyzing the implementation and effectiveness of such methods creates a knowledge gap that limits the adoption of VR in education. Moreover, the absence of rigorous evaluation makes it difficult to identify improvements in the scope of these tools, which is essential for adapting them to the needs of both students and educators. Thus, it is important to explore, document, and analyze usability evaluation methods in VR applications intended for learning. This will contribute to optimizing the design of virtual environments and ensuring a better user experience.

2 Methodology

This research study follows the PRISMA guidelines to ensure the reliability and validity of the results obtained in the systematic review. The methodology is carried out in three stages: identification, screening, and inclusion. In the identification phase, an exhaustive search is conducted in relevant databases, carefully documenting the search terms and criteria used to ensure the reproducibility of the process.

Subsequently, during the screening phase, studies are rigorously selected based on their relevance, methodological quality, and alignment with the research objectives, through a review of titles, abstracts, and full texts. Finally, in the inclusion phase, the selected studies are analyzed in depth, with a more detailed evaluation of their methodological quality in order to strengthen the validity of the findings.

2.1 Eligibility criteria

To refine the selection process, inclusion and exclusion criteria were established and applied across the three phases, as detailed in Table 1

Table 1. Inclusion and Exclusion Criteria of the Study

Inclusion Criteria	**Exclusion Criteria**
Publications from 2020 to 2025	Publications prior to 2020
Publications in English	Publications in languages other than English
Studies addressing usability in virtual reality environments applied to learning	
Studies that clearly describe the methodology used	Studies not related to virtual reality environments applied to learning
Studies that include both quantitative and qualitative results	

2.2. Source of information

The sources of information used in this systematic review were the Scopus and Web of Science (WoS) databases, accessed through the institutional portal of Universidad Señor de Sipán. These platforms were selected due to their extensive coverage of peer-reviewed scientific literature and their relevance in the fields of educational technology, human-computer interaction, and virtual reality. Both databases offer a wide range of high-quality publications, ensuring a rigorous and up-to-date approach in the collection of studies related to usability evaluation in immersive learning environments.

2.3 Search strategy

The bibliometric search process was conducted using the databases provided by Universidad Señor de Sipán, namely Scopus and Web of Science (WOS), on June 20, 2025. It is important to note that searches conducted after this date may yield a greater volume of information due to the ongoing evolution of research in this field. The search equations used, described in Table 2, were carefully designed to ensure the accurate identification of relevant studies on usability evaluation methods in virtual reality and immersive environments.

Table 2. Search strings used

Database	**Search Equation**
SCOPUS	("usability evaluation" OR "usability testing" OR "usability assessment") AND ("virtual reality" OR "VR") AND ("learning applications" OR "educational applications" OR "training systems" OR "learning outcomes OR " user "engagement")
WOS	TS=("usability evaluation" OR "usability testing" OR "usability assessment") AND TS=("virtual reality" OR "VR") AND TS=("learning applications" OR "educational

	applications" OR "training systems" OR "learning outcomes" OR "user engagement")

2.4 Data management

For the management and analysis of the data extracted from the selected databases, tools such as Microsoft Excel and the Python programming language were used. Microsoft Excel was employed in the initial stages to organize and clean the records, facilitating the identification of duplicates, the logging of inclusion and exclusion criteria, and the tracking of the selection process. Subsequently, Python was used to perform more structured and reproducible analyses, including variable coding, bibliographic data processing, and the generation of descriptive statistics. This combination of tools ensured efficient, transparent, and replicable handling of the information collected during the systematic review.

2.5 Selection process

The selection of studies was carried out in multiple stages. First, an initial screening of titles and abstracts was conducted to exclude those that did not meet the objectives of the review. Subsequently, a full-text reading of the preselected studies was performed, assessing their relevance, methodological quality, and thematic alignment. The selection process was conducted in a systematic and well-documented manner, ensuring the traceability of each decision. In cases where doubts or disagreements arose regarding the inclusion of a study, these were resolved through discussion and consensus among the reviewers. To ensure a rigorous selection process aligned with the research objectives, key guiding questions were formulated to steer the systematic review, as shown in Table 3.

Table 3. Research questions posed

No.	Research Questions
RQ1	What methods have been used to evaluate usability in virtual reality environments for educational purposes?
RQ2	What are the most commonly used usability evaluation techniques in studies on virtual reality for learning?
RQ3	What tools, metrics, or protocols are applied to measure usability in these environments?
RQ4	How frequently are recognized standards such as SUS (System Usability Scale), ISO 9241, or other evaluation frameworks used?
RQ5	Are qualitative, quantitative, or mixed methods used to evaluate usability in educational VR contexts?
RQ6	Which objective user performance metrics (interaction time, errors, completion rate) are most frequently used in usability evaluations?

RQ7	How are objective performance metrics related to the subjective perception of usability in immersive educational environments?

3 Results

A total of 393 articles were identified through search strategies in the Web of Science and Scopus databases. After removing 4 duplicates, 389 records were screened. Of these, 218 were deemed irrelevant, and 136 were excluded based on the established criteria. In the end, 29 articles published between January 2020 and June 2025 were included. **Figure 1** presents the PRISMA flow diagram showing the number of articles identified and processed.

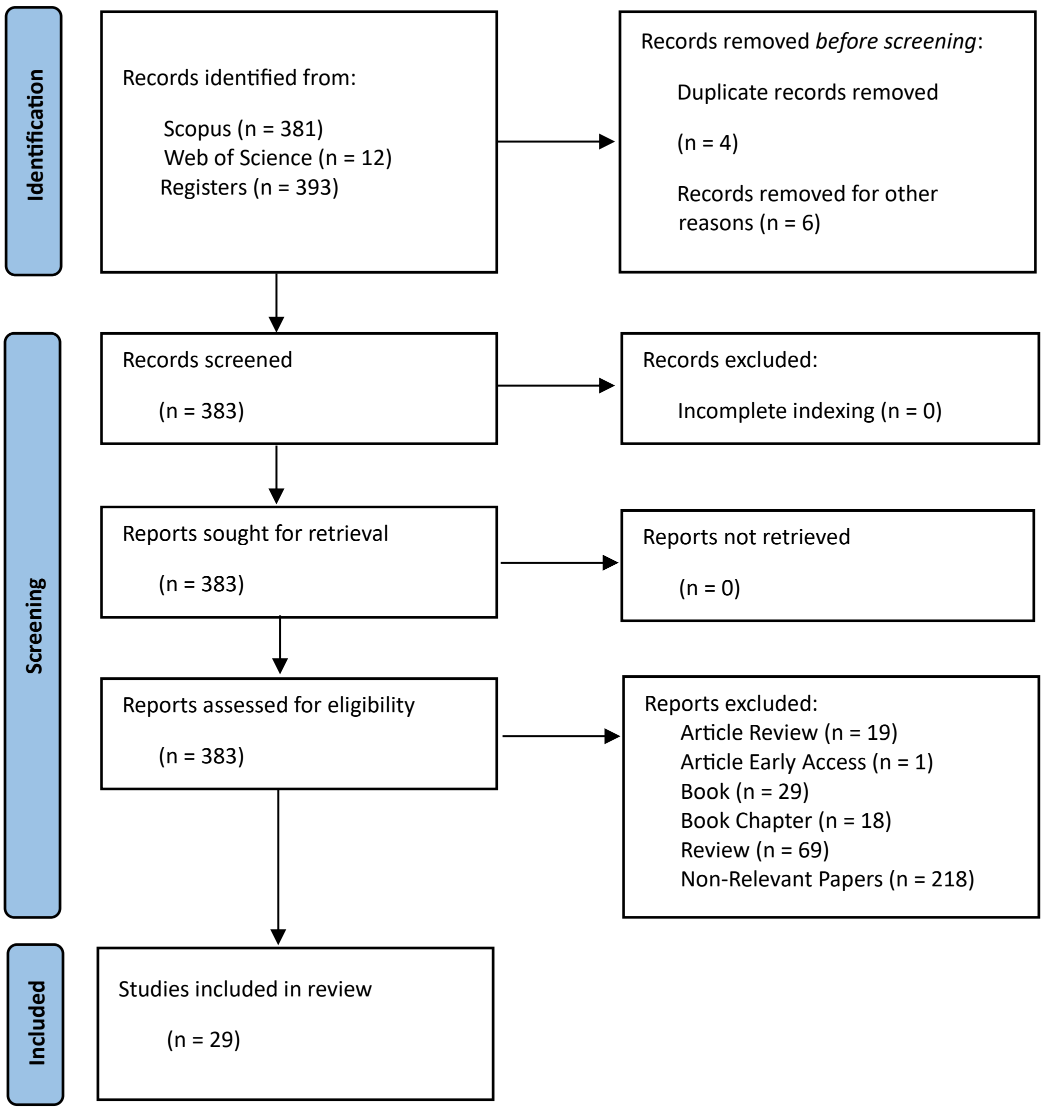

Figure 1. PRISMA flow diagram. Created by the authors based on data from Scopus and WoS.

Table 4 presents a summary table of the articles selected for this systematic review, organized according to key criteria extracted from the methodological analysis. It includes information on the methodological approach used, the evaluation instruments and techniques applied, the context in which they were implemented, and the most relevant findings reported by each study. This systematization allows for the identification of recurring patterns such as the predominance of mixed-methods approaches and the frequent use of standardized questionnaires like SUS or QUX, while also highlighting the diversity of application contexts and outcomes in terms of usability, motivation, and learning. The table serves as a comparative basis for subsequent discussions on trends, methodological gaps, and opportunities for improvement in the evaluation of immersive environments for educational purposes.

Table 4. Summary of Selected Studies

Title	Methodological Approach / Main Instruments	Techniques Applied / Application Context	Key Findings
360°-Based Virtual Field Trips to Waterworks in Higher Education (Wolf et al., 2021)	Mixed (questionnaires, pre/post-test, and interviews). QUX, AEQ, QCM (all validated)	Likert scales, semi-structured interviews, qualitative content analysis, t-tests. Real environment (remote access by students for one week via PC/tablet)	High perceived usability (QUX); positive emotions (AEQ) and high motivation (QCM); improved performance (post-test 92%); positive correlation between perception and performance; suggestions: guided tours, maps, more interactivity
A Large-Scale, Multiplayer Virtual Reality Deployment: A Novel Approach to Distance Education in Human Anatomy (Brown et al., 2023)	Mixed (questionnaires, quizzes, academic performance, and thematic coding). BananaVision, BanAnatomy, Canvas (LMS), pre/post surveys	Likert scales, statistical analysis (t-test and ANOVA), thematic coding of comments. Remote virtual environment (students at home with provided VR equipment); VR and Zoom sessions	VR was as effective as in-person learning, with greater engagement and spatial understanding; positive correlation between 3D confidence and performance; high system satisfaction, though complaints about learning curve and HMD effects
A Method for Evaluating the Learning Concentration in	Mixed: quantitative (concentration, task mastery) and qualitative (emotions,	Facial expression analysis with CNN, visual focus tracking, task metrics,	Concentration directly affects performance, but interface flaws

Head-Mounted Virtual Reality Interaction (Lin et al., 2022)	usage observation). VRESE (VR environment), FERVR (facial recognition), custom cameras	correlations between concentration and performance. Lab: controlled environment with in-house devices (HMD, cameras, VRESE)	reduced results; design improvements increased concentration (+18%) and performance (+15.39%); positive relationship between objective and subjective metrics, with some exceptions
A Preoperative Virtual Reality App for Patients Scheduled for Cardiac Catheterization: Pre-Post Questionnaire Study Examining Feasibility, Usability, and Acceptability (Aardoom et al., 2022)	Mixed: quantitative (SUS, CSQ-8, PQ, ITQ) and qualitative (interviews). Standardized questionnaires (SUS, CSQ-8, PQ, ITQ), phone interviews, Oculus Go	Validated questionnaires for usability, presence, and satisfaction; interviews for perceived effectiveness. Mixed environment: hospital and patients' homes	High usability (SUS: 89.1) and satisfaction (CSQ-8: 27.1); 88% reported better preparation; greater presence and usability from home users; no ISO standards applied, but strong coherence between objective and subjective data
A Transferable Psychological Evaluation of Virtual Reality Applied to Safety Training in Chemical Manufacturing (Poyade et al., 2021)	Mixed (quantitative: SUS, ITC; qualitative: LIWC analysis of verbal feedback). SUS, ITC-Sense of Presence Inventory, ad hoc questionnaire, LIWC	Questionnaires, statistical analysis (Mann-Whitney U), sentiment analysis. Simulated environment based on real plant (chemical industry, GSK Scotland)	VR not inferior to video in learning, but outperformed in presence, confidence, and positive emotions; excellent usability (SUS ~80); highlights psychological and immersive value of VR in industrial contexts
A Virtual Reality Based Gas Assessment Application for Training Gas Engineers (Asghar et al., 2019)	Mixed (SUS + qualitative observation). System Usability Scale (SUS)	SUS questionnaire, task observation in VR. Controlled environment (lab) with supervising engineer	High perceived usability (SUS = 84.06), especially among users under 40; objective metrics like time or errors were not included; recommends integrating such measures in future

			research
Advantages of Virtual Reality Childbirth Education (Siivola et al., 2024)	Mixed (qualitative and quantitative, user-centered design – ISO 9241-210). PSUS (positive SUS version), interviews, observation, learning questionnaires	Cognitive walkthrough, semi-structured interviews, content analysis, pre/post questionnaires, metric correlations. Mixed setting: lab (phase 1) and remote (phase 2) with Oculus Quest 2	High perceived usability (PSUS: 87), increased satisfaction and VR preference; minor technical issues reported; design and accessibility adjustments recommended for broader device compatibility
An Innovative Approach for Online Neuroanatomy and Neurorrehabilitation Teaching Based on 3D Virtual Anatomical Models Using Leap Motion Controller During COVID-19 Pandemic (Obrero-Gaitán et al., 2021)	Mixed (quantitative-qualitative; comprehensive evaluation of educational experience and usability). SUS, GAMEX, SEEQ, DELES, CEVEAPEU, online exams, video recordings	Standardized questionnaires, video rubric scoring, qualitative analysis, subjective/objective metric correlations. Mixed environment (remote with LMC at home + synchronous video support)	High usability (SUS ~80), positive LMC experience (high engagement), strong educational satisfaction; ISO standards not applied but validated tools used; academic performance comparable to control, with greater motivation and self-regulation in experimental group
Assessing the Effectiveness of Virtual Reality Serious Games in Poststroke Rehabilitation: A Novel Evaluation Method (Masmoudi et al., 2024)	Mixed (quantitative-qualitative with innovative emotional recognition approach). MoodMe Framework, DeepFace, Leap Motion, Kinect, in-game scores	Facial emotion recognition, performance analysis, user engagement observation. Mixed environment (lab simulated as rehabilitation center)	Positive emotions correlated with better performance; novel and promising method in clinical settings; SUS and ISO not used; high session completion rate; motivation objectively measured through facial expressions
Countering the Novelty Effect: A Tutorial for Immersive Virtual	Mixed (quantitative-qualitative with focus on user experience). Tcha-Tokey	Likert surveys (engagement, presence, flow, immersion, skill,	High satisfaction in skill and engagement; lower flow scores due to interaction

Reality Learning Environments (Miguel-Alonso et al., 2023)	questionnaire, HTC Vive Pro Eye, Unreal Engine	cybersickness), open comments, statistical analysis. Controlled academic lab environment with iVR equipment	difficulties; cybersickness negatively impacted experience; tutorial helped reduce novelty effect and familiarize users
Data Collection Framework for Context-Aware Virtual Reality Application Development in Unity: Case of Avatar Embodiment (Moon et al., 2022)	Mixed (qualitative interviews + quantitative performance metrics). Custom code, Unity Profiler API (partial), screen recordings	Structured interviews, CPU/memory/FPS measurements. Lab (university, developers)	Framework perceived as easy to use and extensible, though complex functions increased CPU usage; proprietary tools used to avoid system overload; good completion rate with acceptable performance
Designing Usability Evaluation Methodology Framework of Augmented Reality Basic Reading Courseware AR BACA SindD for Down Syndrome Learner (Ramli and Zaman, 2011)	Mixed (formative and summative, qualitative and quantitative). Recordings, checklists, adapted questionnaires	Observation, interviews, informal walkthrough, cognitive walkthrough, heuristic evaluation. Schools (real setting with children with Down syndrome)	Adapted methods applied for users with disabilities; ISO 9241-11 used as framework; no detailed quantitative metrics, but strengths identified in ease of use and intuitive interaction
What Variables Are Connected with System Usability and Satisfaction Results from an Educational Virtual Reality Field Trip (Fink et al., 2023)	Mixed (qualitative and quantitative); Semi-structured interviews, surveys (SUS, IPQ, validated scales)	Interviews, questionnaires (SUS, IPQ, satisfaction and cognitive load scales, etc.); Laboratory setting	Explored usability and satisfaction factors in an educational VR experience. Metrics such as ease of use, presence, cognitive load, and satisfaction were analyzed. Some IPQ items were removed for reliability. The approach identified correlations between perception variables

			and system evaluation.
WeChat Mini Program in Laboratory Biosafety Education Among Medical Students at Guangzhou Medical University: A Mixed Method Study of Feasibility and Usability (Li et al., 2024)	Mixed (qualitative and quantitative); Online surveys (SUS and custom questionnaires), semi-structured interviews	Surveys and interviews; Real educational setting (not specified if lab or classroom)	Demonstrated feasibility and good usability of the WeChat program for biosafety education. Variables measured included ease of use, acquired knowledge, and future use intention. Interviews provided complementary insights into user experience.
Virtual Reality Technology in Construction Safety Training: Extended Technology Acceptance Model (Zhang et al., 2022)	Quantitative; Surveys based on the extended TAM	Surveys and interviews; Construction safety training setting	Found that VR acceptance depends on perceived usefulness, ease of use, attitude, and intention to use. Including self-efficacy and enjoyment improved understanding of barriers to adoption in this context.
Virtual Reality in Museums: Does It Promote Visitor Enjoyment and Learning? (Shahab et al., 2023)	Quantitative; Pre- and post-use surveys	Surveys; Real Museum with VR installation	VR use significantly enhanced visitors' perceived enjoyment and learning, suggesting a positive impact of VR in cultural environments.
Virtual Reality Game for Physical and Emotional Rehabilitation of Landmine Victims (Pérez et al., 2022)	Mixed (qualitative and quantitative); SUS, PACES, GEQ, ad-hoc survey	Expert and user evaluations (interviews + surveys); Controlled setting	Participants reported high usability and enjoyment, along with a positive user experience. The tool was validated by healthcare

			professionals and a real user with an amputation.
User VR Experience and Motivation Study in an Immersive 3D Geovisualization Environment Using a Game Engine for Landscape Design Teaching (Carbonell-Carrera et al., 2021)	Quantitative; UX Questionnaire for Immersive Virtual Environments, Intrinsic Motivation Inventory	Standardized questionnaires and numeric scales; Real classroom setting	Multiple dimensions were assessed: immersion, presence, emotion, and motivation. Results showed a positive impact on student engagement and perception of the 3D environment as an educational tool.
Usability Evaluation of the Preoperative ISBAR Identification Situation Background Assessment and Recommendation Desktop Virtual Reality Application: Qualitative Observational Study (Andreasen et al., 2022)	Mixed (qualitative and quantitative); SUS, direct observation, interviews	Observation, interviews, questionnaires; Laboratory setting	Combined direct observation, qualitative interviews, and SUS scale to assess usability. Identified interface strengths and areas for interaction improvement, highlighting the value of a mixed approach to capture multiple experience dimensions.
Usability Evaluation of Multimodal Interactive Virtual Environments for Learners Who Are Blind: An Empirical Investigation (Darin et al., 2022)	Mixed (qualitative and quantitative); CLUE checklist, observation, interviews	Observation, interviews, specialized checklist; Real-world setting	Evaluated the usability of interactive virtual environments for blind learners using observation, interviews, and the CLUE checklist. The mixed approach captured specific aspects of accessibility, effectiveness, and satisfaction tailored to the sensory and cognitive needs of the

			target audience.
Usability and User Experience of an Individualized and Adaptive Game-Based Therapy for Children with Cerebral Visual Impairment (Ben Itzhak et al., 2023)	Mixed (qualitative and quantitative); Relative Enjoyment Scale, direct observation, custom tools	Direct observation, structured interviews (This-or-That, Laddering technique), usage data analysis; Controlled setting	Assessed usability and user experience of adaptive game-based therapy for children with CVI. Mixed methods captured system effectiveness, clarity, and performance, highlighting customization based on the child's visual profile.
The Design Development and Usability of a Virtual Reality Training Application for the Dental Trainees (Asghar et al., 2022)	Quantitative; SUS questionnaire	SUS survey (System Usability Scale); Controlled setting	Usability of a VR dental training app was assessed using SUS, yielding a score of 82.50, indicating high perceived usability among users.
Knowledge Maps as a Complementary Tool to Learn and Teach Surgical Anatomy in Virtual Reality: A Case Study in Dental Implantology (Lúcio et al., 2024)	Mixed (quantitative and qualitative); Surgical anatomy questionnaires, subjective feedback, direct observation	Performance tests, subjective evaluation; Mixed setting (laboratory and real classroom)	Evaluated the use of knowledge maps in a VR environment to teach surgical dental anatomy. Metrics included completion time, questionnaire scores, satisfaction, and workload. Results showed positive usability and acceptance among students and instructors.
Investigating the Usability of a Head-Mounted Display Augmented Reality Device in Elementary School Children	Mixed (quantitative and qualitative); Standardized tutorial, SUS, direct observation, non-validated custom	User testing, behavioral metrics; Controlled setting	Evaluated how elementary students interacted with HMD-type AR devices. Measured emotions, efficiency, and overall usability. Findings

(Lauer et al., 2021)	tools		showed clear preferences for interaction modes and high system acceptance.
Investigating the Impact of Scenario and Interaction Fidelity on Training Experience When Designing Immersive Virtual Reality-Based Construction Safety Training (Luo et al., 2023)	Quantitative; Igroup Presence Questionnaire, Witmer & Singer's Presence Questionnaire (WSPQ), task completion time	Post-session survey, task completion timing; Controlled setting	Analyzed how scenario and interaction fidelity influenced VR training experiences. Usability was indirectly measured through task completion time, and presence was assessed using standardized questionnaires. Findings suggest higher fidelity improves user experience.
Investigating the Efficacy of a Virtual Reality-Based Testing Station of Flexible Manufacturing System: A Usability and Heuristic Evaluation (Hariyanto et al., 2024)	Mixed (qualitative and quantitative); System Usability Scale (SUS), heuristic evaluation (12 Sutcliffe & Gault heuristics)	SUS questionnaire, expert heuristic evaluation; Implicitly controlled setting	Identified usability strengths with a favorable SUS score, but also limitations related to user efficiency and adaptation. Heuristic evaluation highlighted issues such as visual inconsistency and interface manipulation challenges.
Investigating the Effectiveness of Immersive VR Skill Training and Its Link to Physiological Arousal (Radhakrishnan et al., 2023)	Mixed approach; Instruments: NASA-TLX, physiological sensors (EDA, ECG), adapted questionnaires, performance metrics	Techniques: – Cognitive load assessment (NASA-TLX) – Immersion/presence evaluation (questionnaires) – Physiological measurement (SCR,	VR training was as effective as physical training for skill improvement, but elicited lower physiological arousal and greater enjoyment and immersion.

		HR via Shimmer GSR+ and Polar H10) – Objective performance (TCT, CT) Context: Controlled lab environment with two rooms (VR and physical)	
Interprofessional Team Training With Virtual Reality: Acceptance, Learning Outcome, and Feasibility Evaluation Study (Neher et al., 2024)	Quantitative approach; Validated instruments: SUS, SSQ, NASA-TLX, USEQ, Slater's Presence Scale	Techniques: – Questionnaires: SUS (usability), Slater (presence), NASA-TLX (cognitive load), SSQ (adverse effects), USEQ (satisfaction) – Demographic and confidence data Context: Home-based e-learning + in-person session using Meta Quest 2	Good usability (SUS: 72.5) despite high cognitive load (NASA-TLX: 64.5). Technical issues did not hinder task completion. Significant improvement in handover skills (HAT) after training.
Improving Ray Tracing Understanding With Immersive Environments (Trindade et al., 2024)	Mixed (quantitative and qualitative); Instruments: usability tests, validated SUS questionnaire, non-validated feedback survey, non-validated theoretical test	Techniques: – Usability tests with predefined tasks – SUS questionnaire at the end – Qualitative feedback survey – Pre- and post-theoretical test Context: Controlled lab setting (INESC-ID) with Oculus Rift	– Average SUS score: 78.8 ("good" usability) – 46.27% improvement in conceptual understanding – Positive feedback on immersive experience – Improvement areas identified through participant comments

Figure 2 shows the distribution of articles by year of publication, allowing for the identification of certain trends in scientific production related to user experience evaluation in educational virtual reality applications. The year 2020 had the lowest number of publications found, possibly influenced by the global context of the pandemic. In contrast, 2021 and 2022 saw a significant increase in the number of studies, representing the peak of research interest. However, from 2023 and 2024 onward, there is a slight decline in output, which may reflect a shift in research priorities or a diversification of approaches in emerging technologies.

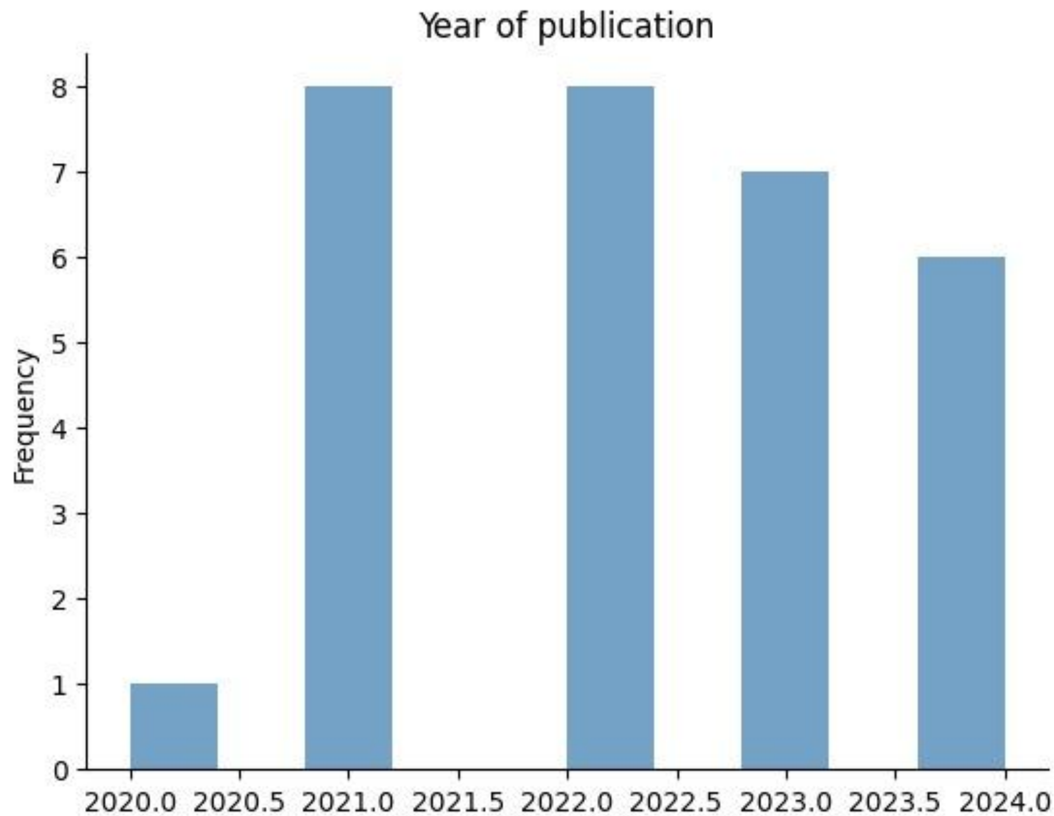

Figure 2. Distribution of Included Articles by Year of Publication

Regarding RQ1, most studies employed a mixed-method approach to usability evaluation, combining qualitative techniques such as observation and interviews with quantitative methods such as standardized Likert-scale-based questionnaires. This predominance of the mixed-method approach responds to the need to obtain both subjective user experience data and numerical evidence that enables comparisons and statistical analysis. For example, the articles by Pérez et al. (2022) and Fink et al. (2023) describe the implementation of user testing protocols in which, after using the VR application, users' perceptions were collected through Likert-type questionnaires, followed by interviews to explore aspects such as satisfaction, difficulties, or suggestions for improvement. However, the study by Zhang et al. (2022) adopted an alternative approach based on the Technology Acceptance Model (TAM) to evaluate the acceptance of VR technology in a construction training context, motivated by the previously low adoption of VR in that sector.

In relation to RQ2, the vast majority of studies relied on a combination of surveys, interviews, and standardized questionnaires as the main means of collecting user experience data (Andreasen et al., 2022; Fink et al., 2023; Li et al., 2024). These techniques allow, on the one hand, for the collection of quantitative measurements (e.g., ratings using Likert scales or validated instruments) and, on the other hand, for a deeper exploration of user perceptions, motivations, and barriers through semi-structured or focused interviews. This predominance reflects the need to capture both the extent of certain usability aspects and their underlying causes.

However, some studies went beyond these mentioned methods. The article by Darin et al. (2022), for instance, combined surveys and interviews with direct observation of user interaction in the VR environment and a specialized checklist. Specifically, observation allowed the research team to identify behaviors or errors that participants may not have spontaneously mentioned during interviews, such as imprecise gestures or inefficient exploration patterns. The checklist, in turn, was designed based on usability criteria adapted to immersive environments.

In contrast, in a small number of cases (Shahab et al., 2023; Asghar et al., 2022), the studies relied exclusively on either interviews or purely quantitative surveys. However, these approaches may present limitations in capturing the full scope of the user experience.

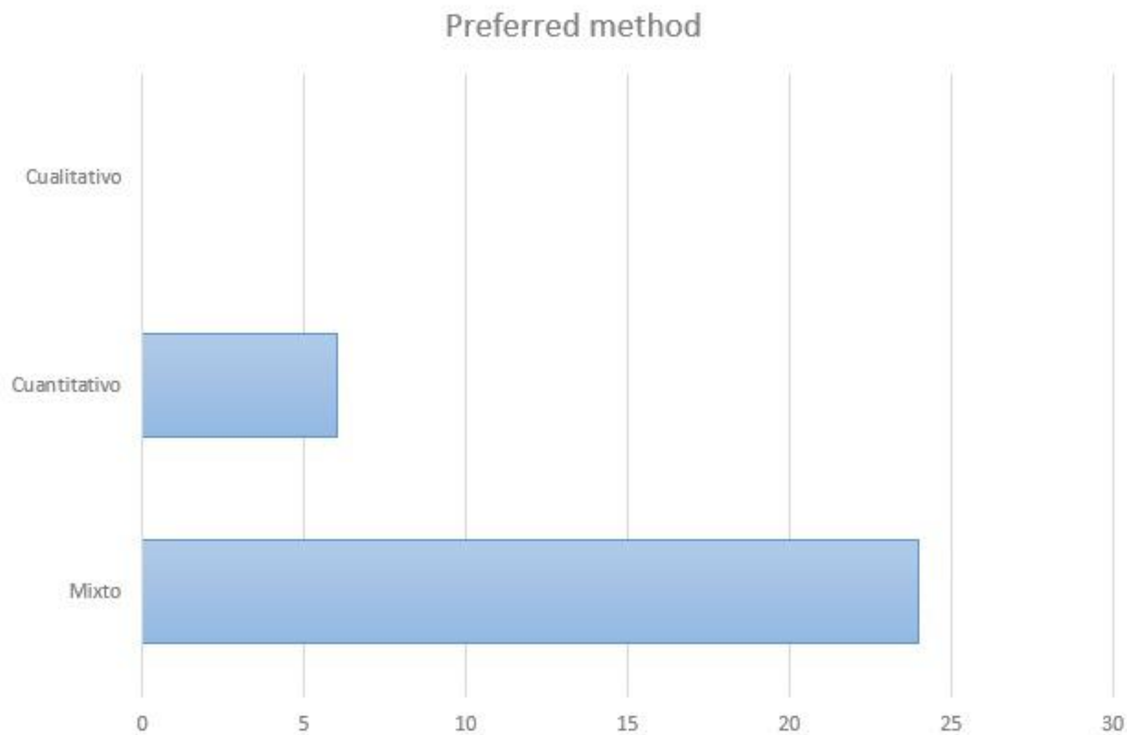

Figure 3. Distribution of Approaches and Methods Used in the Articles.

Respecto a la PI3. Para medir la usabilidad múltiples estudios (Wolf et al., 2021), (Poyade et al., 2021) (Asghar et al., 2019) en su mayoría recurre a la System Usability Scale (SUS) para obtener una valoración global, mientras que el NASA-TLX se utiliza en (Radhakrishnan et al., 2023) (Neher et al., 2024) para cuantificar la carga cognitiva durante las tareas. Para evaluar la sensación de presencia e inmersión, se aplican tanto la escala de Slater como el Igroup Presence Questionnaire (IPQ) en (Fink et al., 2023)(Neher et al., 2024). Además, muchos trabajos complementan estos datos cuantitativos con encuestas de feedback cualitativo y, en algunos casos (Ramli and Zaman, 2011), (Moon et al., 2022) con grabaciones de pantalla, listas de verificación y cuestionarios adaptados, lo que permite identificar con más detalle los puntos de fricción y las oportunidades de mejora. Así, las métricas más habituales son la usabilidad del sistema, la carga cognitiva, la presencia e immersión y la satisfacción o facilidad de uso, destacando especialmente el uso del SUS como referente principal.

Regarding PI3, to assess usability, multiple studies (Wolf et al., 2021; Poyade et al., 2021; Asghar et al., 2019) primarily rely on the System Usability Scale (SUS) to obtain a general usability score, while NASA-TLX is used in (Radhakrishnan et al., 2023; Neher et al., 2024) to quantify the cognitive load during tasks. To evaluate the sense of presence and immersion, both the Slater scale and the Igroup Presence Questionnaire (IPQ) are applied in (Fink et al., 2023; Neher et al., 2024). In addition, many studies complement these quantitative data with qualitative feedback surveys, and in some cases (Ramli and Zaman, 2011; Moon et al., 2022), with screen recordings, checklists, and custom questionnaires, allowing for a more detailed identification of friction points and improvement opportunities. Therefore, the most common metrics are system usability, cognitive load, presence and immersion, and user satisfaction or ease of use, with SUS standing out as the primary reference.

Regarding PI4, out of the 30 reviewed studies, 47% used the SUS to evaluate usability, while the other 16 did not use this questionnaire. Only two studies employed ISO 9241: (Siivola et al., 2024) applied ISO 9241-210, and (Ramli and Zaman, 2011) was based on ISO 9241-11. Additionally, two studies (Neher et al., 2024; Zhang et al., 2022; Ramli and Zaman, 2011) used the Technology Acceptance Model (TAM) as an alternative analytical framework, and in one case (Ramli and Zaman, 2011), the Common Industry Format (CIF) for usability testing was employed.

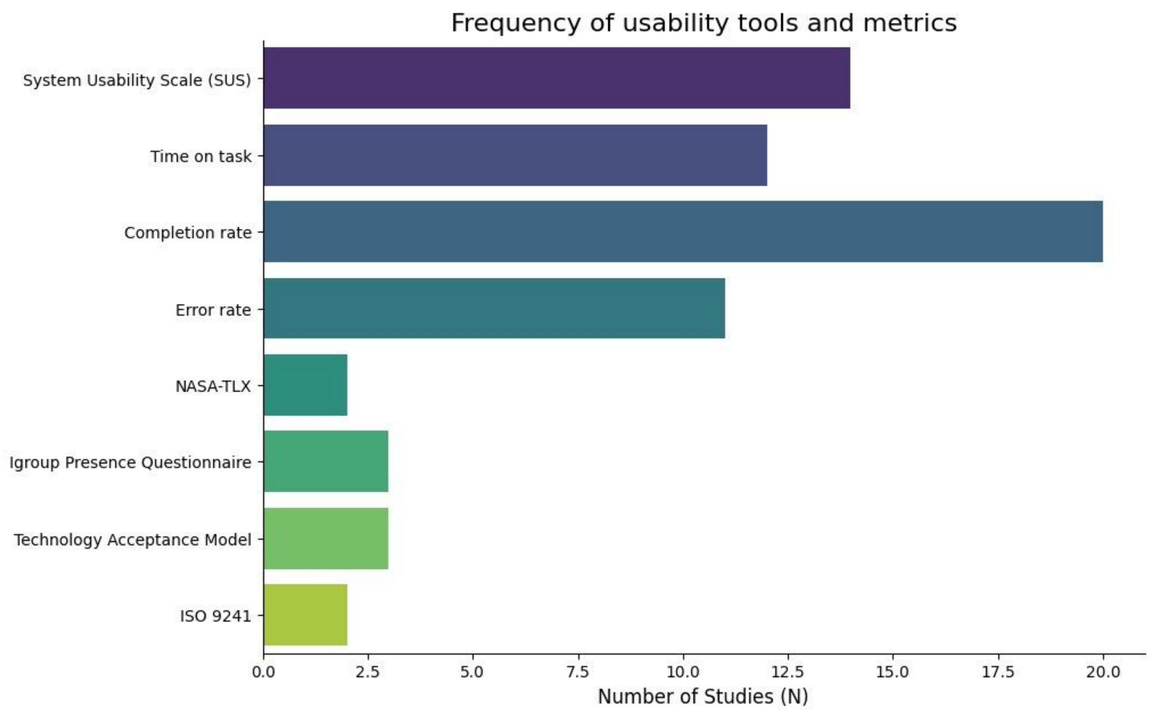

Figure 4. Frequency of Use of Usability Tools and Metrics

Regarding PI 5, the vast majority of studies 24 out of 30, or 80% employ a mixed-methods approach, combining quantitative measurements such as standardized questionnaires (SUS, QUX, AEQ, QCM), pre-/post-knowledge tests, and objective metrics, with qualitative methods like semi-structured interviews, open-ended questions, focus groups, screen recordings, and content analysis of responses. Only 20% of the studies relied solely on quantitative methods, primarily based on Likert-scale questionnaires, exam scores, and retention statistics, and none used exclusively qualitative techniques. This predominance of mixed-methods allows researchers to capture both objective data on performance and cognitive load, as well as user perceptions, emotions, and suggestions for improvement.

As for PI 6, among the most frequently used objective metrics are task completion time measured in 42% of the studies (Wolf et al., 2021; Siivola et al., 2024; Miguel-Alonso et al., 2023) error tracking during interaction, reported in 11 studies, and task success or completion rate, assessed in 20 investigations. Around one-third of the studies conducted post-training exams to measure knowledge retention. Other less common indicators include improvement score composites (IS), skill transfer metrics (HAT), and gameplay data analysis, which provide complementary insights into users' objective performance.

Regarding PI 7, only a small portion of the reviewed articles approximately 12 out of 30 explicitly contrast objective performance data with subjective perceptions of usability (Wolf et al., 2021; Brown et al., 2023). In those cases, the general trend points to a moderate positive correlation: high scores on scales like SUS, QUX, or CSQ-8 often accompany faster task times, fewer errors, or higher post-test gains (Wolf et al., 2021). Similarly, students who reported higher spatial confidence achieved better quiz and exam results (Brown et al., 2023). However, there are subtle discrepancies some studies found specific inconsistencies, such as participants with high objective focus but low scores due to interface issues (Lin et al., 2022), or high levels of cybersickness that drastically reduced satisfaction despite successful task completion (Miguel-Alonso et al., 2023). Physiological effects linked to improved performance were also observed, which were not always reflected in subjective questionnaires (Siivola et al., 2024; Trindade et al., 2024; Neher et al., 2024). Overall, the evidence suggests that while usability perceptions generally align with measured performance, factors such as interactive design quality, added cognitive load, or physical discomfort can break this correspondence highlighting the importance of combining evaluation methods.

4 Discussion

The results of this systematic review reveal clear trends and methodological gaps in the use of virtual reality (VR) for educational purposes, particularly in the evaluation of user experience (UX) and usability. Mixed-methods approaches prevail, combining standardized questionnaires, such as Likert scales, with interviews and direct observation. This approach aims to capture both objective performance indicators and subjective perceptions and barriers during interaction in immersive environments. The System Usability Scale (SUS) stands out as the most widely used tool (47% of the studies), supporting its usefulness as an accessible and validated metric. However, there is limited adoption of more robust international standards, such as ISO 9241 or the Common Industry Format (CIF), which constrains the comparability and rigor of the studies.

Although many works combine objective data such as task completion time, error rate, or knowledge retention with subjective metrics, only around 40% of the studies explicitly contrasted both approaches. In those cases, a moderate correlation was found between high perceived usability and better performance. However, discrepancies were also reported: symptoms of cybersickness, cognitive overload, or deficiencies in interactive design negatively affected the experience, despite positive outcomes in some metrics. Additionally, most studies focus on simple or short-duration tasks, with few addressing complex structures or evaluating the transfer of skills to real-world contexts thus limiting our understanding of VR's long-term educational impact. Lastly, alternative frameworks such as the Technology Acceptance Model (TAM) or immersion and presence scales (e.g., IPQ, Slater) offer valuable perspectives but remain underutilized and require more systematic integration into evaluation protocols.

5 Conclussions

This systematic review reveals that the evaluation of user experience in educational virtual reality applications has evolved toward mixed-methods approaches, integrating both quantitative and qualitative methodologies. Instruments such as the SUS, NASA-TLX, and immersion scales emerge as recurrent tools to assess usability, cognitive load, and the quality of immersive experience. However, persistent limitations are identified, including the limited application of international standards, the low frequency of longitudinal studies, and the limited inclusion of control groups. Moreover, the relationship between subjective perception and objective performance is influenced by factors such as interface ergonomics, cognitive load, or physical discomfort, highlighting the need for comprehensive and contextualized assessments.

It is recommended that future research develop more rigorous and standardized protocols, combining subjective and objective metrics such as physiological data or interaction patterns, and addressing more complex tasks that enable the evaluation of skill transfer to real-world contexts. Likewise, the adoption of established normative frameworks would allow for more precise and reproducible comparisons. Taken together, these findings provide a critical and up-to-date perspective on evaluation practices in immersive educational environments, laying the groundwork for research that enhances user experience, usability, and the pedagogical impact of virtual reality.

References

Aardoom, J. J., Hilt, A. D., Woudenberg, T., Chavannes, N. H., and Atsma, D. E. (2022). A Preoperative Virtual Reality App for Patients Scheduled for Cardiac Catheterization: Pre-Post Questionnaire Study Examining Feasibility, Usability, and Acceptability. *JMIR Cardio* 6. doi: 10.2196/29473,

Andreasen, E. M., Høigaard, R., Berg, H., Steinsbekk, A., and Haraldstad, K. (2022). Usability Evaluation of the Preoperative ISBAR (Identification, Situation, Background, Assessment, and Recommendation) Desktop Virtual Reality Application: Qualitative Observational Study. *JMIR Hum Factors* 9. doi: 10.2196/40400,

Asghar, I., Egaji, O. A., Dando, L., Griffiths, M., and Jenkins, P. (2019). A virtual reality based gas assessment application for training gas engineers. *ACM International Conference Proceeding Series*, 57–61. doi: 10.1145/3357419.3357443;SUBPAGE:STRING:ABSTRACT;CSUBTYPE:STRING:CONFERENCE

Asghar, I., Griffiths, M., Dando, L., and Salisbury, I. (2022). The Design, Development and Usability of a Virtual Reality Training Application for the Dental Trainees. *ACM International Conference Proceeding Series*. doi: 10.1145/3549865.3549881

Ben Itzhak, N., Franki, I., Jansen, B., Kostkova, K., Wagemans, J., and Ortibus, E. (2023). Usability and user experience of an individualized and adaptive game-based therapy for children with cerebral visual impairment. *Int J Child Comput Interact* 35, 100551. doi: 10.1016/J.IJCCI.2022.100551

Brown, K. E., Heise, N., Eitel, C. M., Nelson, J., Garbe, B. A., Meyer, C. A., et al. (2023). A Large-Scale, Multiplayer Virtual Reality Deployment: A Novel Approach to Distance Education in Human Anatomy. *Med Sci Educ* 33, 409–421. doi: 10.1007/S40670-023-01751-W, Carbonell-Carrera, C., Saorin, J. L., and Díaz, D. M. (2021). User VR Experience and Motivation Study in an Immersive 3D Geovisualization Environment Using a Game Engine for Landscape Design Teaching. *Land 2021, Vol. 10, Page 492* 10, 492. doi: 10.3390/LAND10050492

Darin, T., Andrade, R., and Sánchez, J. (2022). Usability evaluation of multimodal interactive virtual environments for learners who are blind: An empirical investigation. *Int J Hum Comput Stud* 158, 102732. doi: 10.1016/J.IJHCS.2021.102732

Fink, M. C., Eisenlauer, V., and Ertl, B. (2023). What variables are connected with system usability and satisfaction? Results from an educational virtual reality field trip. *Computers & Education: X Reality* 3, 100043. doi: 10.1016/J.CEXR.2023.100043

Hariyanto, D., Hakim, V. G. Al, Husna, A. F., Badarudin, R., Yuniarti, N., and Adinda, D. (2024). Investigating the Efficacy of a Virtual Reality-Based Testing Station of Flexible Manufacturing System: A Usability and Heuristic Evaluation. *International Journal of Online and Biomedical Engineering (iJOE)* 20, 67–82. doi: 10.3991/IJOE.V20I08.47883

Johnson, D. M. (2005). Introduction to and Review of Simulator Sickness Research. *PsycEXTRA Dataset*. doi: 10.1037/E456932006-001

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *Int J Aviat Psychol* 3, 203–220. doi: 10.1207/S15327108IJAP0303_3;WEBSITE:WEBSITE:TFOPB;PAGEGROUP:STRING:PUBLICATION

Lauer, L., Altmeyer, K., Malone, S., Barz, M., Brünken, R., Sonntag, D., et al. (2021). Investigating the usability of a head-mounted display augmented reality device in elementary school children. *Sensors* 21, 6623. doi: 10.3390/S21196623/S1

Li, Q. J., Zhao, J. J., Yan, R. C., Gao, Q. J., Zhen, Y., Li, X., et al. (2024). WeChat mini program in laboratory biosafety education among medical students at Guangzhou Medical University: a mixed method study of feasibility and usability. *BMC Med Educ* 24, 1–11. doi: 10.1186/S12909-024-05131-9/TABLES/4

Lin, Y., Lan, Y., and Wang, S. (2022). A method for evaluating the learning concentration in head-mounted virtual reality interaction. *Virtual Real* 27, 863–885. doi: 10.1007/S10055-022-00689-5

Lúcio, I. M., de Faria, B. G., Raidou, R. G., Proença, L., Zagalo, C., Mendes, J. J., et al. (2024). Knowledge maps as a complementary tool to learn and teach surgical anatomy in virtual reality: A case study in dental implantology. *Healthc Technol Lett* 11, 289–300. doi: 10.1049/HTL2.12094,

Luo, Y., Ahn, S., Abbas, A., Seo, J. O., Cha, S. H., and Kim, J. I. (2023). Investigating the impact of scenario and interaction fidelity on training experience when designing immersive virtual reality-based construction safety training. *Developments in the Built Environment* 16, 100223. doi: 10.1016/J.DIBE.2023.100223

Makransky, G., and Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development* 66, 1141–1164. doi: 10.1007/S11423-018-9581-2/METRICS

Masmoudi, M., Zenati, N., Izountar, Y., Benbelkacem, S., Haicheur, W., Guerroudji, M. A., et al. (2024). Assessing the effectiveness of virtual reality serious games in post-stroke rehabilitation: a novel evaluation method. *Multimed Tools Appl* 83, 36175–36202. doi: 10.1007/S11042-023-17980-5/METRICS

Maulidiya, D., Nugroho, B., Santoso, H. B., and Hasibuan, Z. A. (2024). Thematic evolution of smart learning environments, insights and directions from a 20-year research milestones: A bibliometric analysis. *Heliyon* 10, e26191. doi: 10.1016/J.HELIYON.2024.E26191/ASSET/8D96B7AD-2536-4AC4-9BA0-A62ACD66D56B/MAIN.ASSETS/GR1.JPG

Miguel-Alonso, I., Rodriguez-Garcia, B., Checa, D., and Bustillo, A. (2023). Countering the Novelty Effect: A Tutorial for Immersive Virtual Reality Learning Environments. *Applied Sciences (Switzerland)* 13, 593. doi: 10.3390/APP13010593/S1

Moon, J., Jeong, M., Oh, S., Laine, T. H., and Seo, J. (2022). Data Collection Framework for Context-Aware Virtual Reality Application Development in Unity: Case of Avatar Embodiment. *Sensors 2022, Vol. 22, Page 4623* 22, 4623. doi: 10.3390/S22124623

Neher, A. N., Wespi, R., Rapphold, B. D., Sauter, T. C., Kämmer, J. E., and Birrenbach, T. (2024). Interprofessional Team Training With Virtual Reality: Acceptance, Learning Outcome, and Feasibility Evaluation Study. *JMIR Serious Games* 12. doi: 10.2196/57117,

Obrero-Gaitán, E., Nieto-Escamez, F. A., Zagalaz-Anula, N., and Cortés-Pérez, I. (2021). An Innovative Approach for Online Neuroanatomy and Neurorrehabilitation Teaching Based on 3D Virtual Anatomical Models Using Leap Motion Controller During COVID-19 Pandemic. *Front Psychol* 12. doi: 10.3389/FPSYG.2021.590196,

Papaioannou, G., Volakaki, M.-G., Kokolakis, S., and Vouyioukas, D. (2023). Learning Spaces in Higher Education: A State-of-the-Art Review. *Trends in Higher Education 2023, Vol. 2, Pages 526-545* 2, 526–545. doi: 10.3390/HIGHEREDU2030032

Paz, F., Zapata, C., Olivares, C., Apaza, S., and Pow-Sang, J. A. (2015). Usability evaluation of educational tools: A systematic review. 27–38. Available at: https://cris.pucp.edu.pe/es/publications/usability-evaluation-of-educational-tools-a-systematic-review (Accessed July 21, 2025).

Pérez, V. Z., Yepes, J. C., Vargas, J. F., Franco, J. C., Escobar, N. I., Betancur, L., et al. (2022). Virtual Reality Game for Physical and Emotional Rehabilitation of Landmine Victims. *Sensors* 22, 5602. doi: 10.3390/S22155602/S1

Poyade, M., Eaglesham, C., Trench, J., and Reid, M. (2021). A Transferable Psychological Evaluation of Virtual Reality Applied to Safety Training in Chemical Manufacturing. *ACS Chemical Health and Safety* 28, 55–65. doi: 10.1021/ACS.CHAS.0C00105/SUPPL_FILE/HS0C00105_SI_003.XLSX

Radhakrishnan, U., Chinello, F., and Koumaditis, K. (2023). Investigating the effectiveness of immersive VR skill training and its link to physiological arousal. *Virtual Real* 27, 1091–1115. doi: 10.1007/S10055-022-00699-3,

Ramli, R., and Zaman, H. B. (2011). Designing usability evaluation methodology framework of Augmented Reality basic reading courseware (AR BACA SindD) for Down Syndrome learner. *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, ICEEI 2011*. doi: 10.1109/ICEEI.2011.6021807

Shahab, H., Mohtar, M., Ghazali, E., Rauschnabel, P. A., and Geipel, A. (2023). Virtual Reality in Museums: Does It Promote Visitor Enjoyment and Learning? *Int J Hum Comput Interact* 39, 3586–3603. doi: 10.1080/10447318.2022.2099399

Siivola, M., Leinonen, T., and Malmi, L. (2024). Advantages of virtual reality childbirth education. *Computers & Education: X Reality* 4, 100058. doi: 10.1016/J.CEXR.2024.100058

Sulisworo, D., Erviana, V. Y., and Robi'in, B. (2023). Virtual Reality in Education Designing Immersive and Innovative Learning Experiences.

Sutcliffe, A., and Gault, B. (2004). Heuristic evaluation of virtual reality applications. *Interact Comput* 16, 831–849. doi: 10.1016/J.INTCOM.2004.05.001

Trindade, N. V., Custódio, L., Ferreira, A., and Pereira, J. M. (2024). Improving Ray Tracing Understanding With Immersive Environments. *IEEE Transactions on Learning Technologies* 17, 1975–1988. doi: 10.1109/TLT.2024.3436656

Wolf, M., Wehking, F., Montag, M., and Söbke, H. (2021). 360°-Based Virtual Field Trips to Waterworks in Higher Education. *Computers 2021, Vol. 10, Page 118* 10, 118. doi: 10.3390/COMPUTERS10090118

Zhang, M., Shu, L., Luo, X., Yuan, M., and Zheng, X. (2022). Virtual reality technology in construction safety training: Extended technology acceptance model. *Autom Constr* 135, 104113. doi: 10.1016/J.AUTCON.2021.104113