Evaluating healthcare learning in virtual reality: a systematic review

Vitor dos Santos Silva
Escola Politécnica
Universidade de São Paulo
São Paulo, Brazil
ORCID 0000-0002-6939-3696

Fátima L. S. Nunes
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
São Paulo, Brazil
ORCID 0000-0003-0040-0752

***Abstract*—Virtual Reality (VR) is increasingly employed in healthcare training because it can create interactive three-dimensional learning environments that minimize risk to patients. As VR-based educational interventions become more widespread, evaluating their learning outcomes is becoming a critical focus for researchers and practitioners alike. Although there are a few other studies which verified VR learning in healthcare, they had as their subject the effectiveness of VR itself, not the methods which are used to evaluate learning. This systematic review examined 72 studies (selected from an initial pool of 22773 retrieved from IEEE Xplore, Scopus, Springer, ACM, and PubMed) that span a range of healthcare domains, with a primary focus on surgery and rehabilitation. The studies demonstrated a broad diversity of evaluation methods, including objective measures, subjective questionnaires, and AI-assisted assessments. While the overall methodological quality was generally high, the significant variability in evaluation approaches creates challenges for cross-study comparison and synthesis. The findings underline the necessity for validated, VR-specific assessment frameworks to facilitate the development and expansion of VR applications in lesser-explored areas of healthcare education.**

***Index Terms*—systematic review, virtual reality, learning, evaluation, healthcare**

I. Introduction

In recent years, Virtual Reality (VR) has emerged as a highly promising technology, being adopted in many different fields such as education, healthcare, entertainment, and beyond. Among its many applications, VR has gained recognition as an effective tool for training students and professionals in healthcare. The immersive 3D environments provided by VR enable users to experience scenarios with a depth of presence that is not easy to attain through conventional computer-based simulators. Moreover, compared to real-life, in-field training, VR solutions are often both more cost-effective and safer, since they eliminate direct risks to patients [1].

Using Virtual Reality, trainees can practice complex medical procedures and engage with high-pressure scenarios within a safe, controlled environment that poses no risk to actual patients. In addition, this approach encourages active learning by offering real-time feedback and opportunities for repeated practice, both of which are well-known to be highly beneficial for effective learning [2]. These features collectively position VR as a powerful tool for training in critical contexts such as surgery, intensive care units (ICU), and rehabilitation settings.

Nonetheless, as the adoption of VR in education continues to expand, so does the growing need to rigorously assess the efficacy of various VR-based teaching techniques and methods. Evaluating learning in VR environments can be accomplished through a range of techniques, from written questionnaires to the collection of objective and qualitative metrics during the training session itself [3]. This diversity highlights not just the multifaceted nature of the learning process but also a current lack of consensus about which evaluation methods are most suitable for specific situations.

Numerous studies have demonstrated the potential advantages of this innovative technology for training healthcare professionals, especially when compared with more traditional educational methods [4, 5]. Other studies, such as [6, 7, 8, 9, 10, 11], have evaluated learning in VR, but their main subject was the impact on the users, while this review has as its subject the evaluation methods themselves. However, the wide amount of evaluation methods complicates efforts to compare findings across different studies or to synthesize their results into generalizable insights. Furthermore, whereas some methods emphasize objective outcomes (such as accuracy or time required to complete a task), other approaches focus on subjective aspects, including participants' self-confidence, perceived realism of the simulation, or satisfaction with the learning experience [12].

Given this context, the main objective of the present work is to carry out a systematic review of the different methods and techniques currently used to evaluate learning outcomes in Virtual Reality training programs within the healthcare sector. The goal is to explore, compare, and critically analyze these approaches in order to understand their effectiveness, advantages, and limitations in different educational contexts. To achieve this, the review addresses the following research question:

- What are the existing methods for evaluating learning outcomes in virtual reality-based training within the healthcare field?

II. Theoretical background

Evaluation is a key element in understanding the effectiveness of immersive virtual reality environments for specific applications, such as healthcare learning. To evaluate means to identify the extent to which the developed application and the technology itself contribute to learning. In this sense, a number of methods and metrics have been explored in recent years to provide a reliable analysis [13].

Generally speaking, evaluation methods can be divided into different categories, depending on the aspects considered. One of the categories used in this review refers to the subjectivity of evaluation, dividing methods into objective and subjective. Objective evaluation refers to methods that rely on quantifiable data collected during system usage. Heart rate or electroencephalographic activity are two examples, as well as the force applied on haptic devices [14]. This type of approach, although still subject to some bias, tends to minimize it.

Subjective evaluation, however, focuses on the users' perception of the activity or of their learning. This usually involves questionnaires applied after usage, with questions presented on a scale. The questions often address aspects such as usability, presence, cognitive load, and user satisfaction [15, 16]. Although these measures depend on individual perception, they capture characteristics that cannot be assessed objectively, and are therefore also important.

Another category that can be used to classify evaluation methods is whether they are statistical or AI-based. More recently, AI-based methods have started to gain relevance in the virtual reality community [17], which is particularly useful when dealing with large amounts of data. These techniques make it possible to identify specific patterns that differentiate new users from experienced ones, predict future performance, and even estimate cognitive load. In healthcare, this can be especially valuable for identifying weaker areas in learning [18].

Statistical methods, on the other hand, rely on more classical approaches, such as hypothesis testing, variance analysis (ANOVA), and correlation analysis to validate differences between groups, such as experimental and control groups [19]. Techniques such as Bayesian methods or multivariate analysis are generally more explainable and allow researchers to identify and understand complex relationships between variables [20].

In terms of the elements usually evaluated, the state of the art points to several interrelated dimensions. From a cognitive perspective, elements such as learning, knowledge retention, and decision-making are commonly evaluated [21]. From a psychomotor perspective, measures include movement precision, response time, and motion coordination [22]. Furthermore, aspects related to the quality of the virtual reality application itself are also considered, such as graphic quality, latency, interaction fluency, and overall user experience [23].

III. Protocol

A. Search strategy

This systematic review was carried out using five leading digital databases: IEEE Xplore, Scopus, Springer, ACM, and PubMed, which were consulted on June 2025. These databases were chosen by visually inspecting the first 50 results returned by Google Scholar when queried with the search string, and the analysis was limited to these five databases, as they were considered comprehensive enough for this subject. The inclusion was further limited to articles published in peer-reviewed journals to ensure quality and reliability. The search strategy focused on four key conceptual categories, each defined by a set of relevant keywords, as summarized in Table I.

TABLE I
KEYWORDS OF SEARCH STRING

Category	Keywords
Evaluation	assessment, educational outcomes, evaluation, knowledge retention, learning assessment, learning evaluation, measurement
Virtual Reality	augmented reality, extended reality, simulation, simulator, virtual reality
Health	clinical, health, healthcare, medical, medical procedure, medical training, surgery, surgical, surgical skills
Training and Education	immersive education, immersive learning, training

These keywords were combined into a comprehensive search string as follows: ("augmented reality" OR "extended reality" OR "simulation" OR "simulator" OR "virtual reality") AND ("clinical" OR "health" OR "healthcare" OR "medical" OR "medical procedure" OR "medical training" OR "surgery" OR "surgical" OR "surgical skills") AND ("immersive education" OR "immersive learning" OR "training") AND ("assessment" OR "educational outcomes" OR "evaluation" OR "knowledge retention" OR "learning assessment" OR "learning evaluation" OR "measurement").

B. Inclusion and exclusion criteria

The systematic review adhered to the following inclusion criteria:

- The study must use technologies related to virtual reality, extended reality, or augmented reality.
- The research must be situated within the medical domain.
- Each study must provide a clear description of the learning evaluation methods employed.

The exclusion criteria were defined as follows:

- Studies that employed non-interactive 3D environments.
- Studies with fewer than 10 pages, or where the number of pages could not be determined from the metadata.
- Studies that did not match the following search string (in either their title or abstract): ("virtual reality") AND ("clinical" OR "health" OR "healthcare" OR "medical" OR "medical procedure" OR "medical training" OR "surgery" OR "surgical" OR "surgical skills") AND ("training") AND ("knowledge retention" OR "evaluation") AND NOT "review".

A slightly more restrictive search string was applied when assessing exclusion criteria. This was necessary to ensure that only studies most closely related to the specific research question were retained for review. The term "review" was actively excluded to filter out other systematic reviews from the sample. To standardize the search process, regular expressions implemented in Python were used, compensating for differences in how each database engine processes queries.

C. Initial selection

During the initial selection phase, articles were considered if the search terms appeared in either the title or abstract. Subsequently, the titles and abstracts of these articles underwent further screening to validate compliance with all inclusion and exclusion criteria. Only those meeting all requirements were advanced to the full-text review for final inclusion in the study.

D. Quality evaluation

To assess the quality of included studies, a set of questions, summarized in Table II, was used to rate methodological rigor.

TABLE II
QUALITY ASSESSMENT CRITERIA

Criteria	Value
Does the study have a quantitative method of learning assessment?	Yes (1) or No (0)
Does the study experiment with users?	Yes, with final user (1) or Yes, with an intermediate user (0.5) or No (0)
Are the methods described enough to ensure reproducibility?	Yes (1) or Partially (0.5) or No (0)
Did the study compare its results to the state of the art?	Yes (1) or Partially (0.5) or No (0)

Each question was equally weighted, and studies were awarded a score between 0 and 1 per criterion. The overall quality score for each article thus ranged from 0 to 4. The quality of the majority of studies was high, as will be presented in Section IV-B, and thus it was not used as an exclusion criterion.

IV. RESULTS AND DISCUSSION

The initial search yielded 22773 papers, from which 72 satisfied all inclusion and exclusion criteria after thorough screening. An overview of the selection process can be found in Figure 1.

A. Characteristics of the selected studies

The final sample consisted of studies published from 2003 through 2025. The amount of works published by year can be seen in Figure 2. It highlights the growing relevance of this field in recent years. These works employed a wide range of VR and AR technologies, including hardware such as the HTC Vive or Microsoft HoloLens, which will be further detailed in upcoming sections. Every included study utilized at least one objective evaluation technique, and many combined both objective and subjective methods in their assessment of learning outcomes.

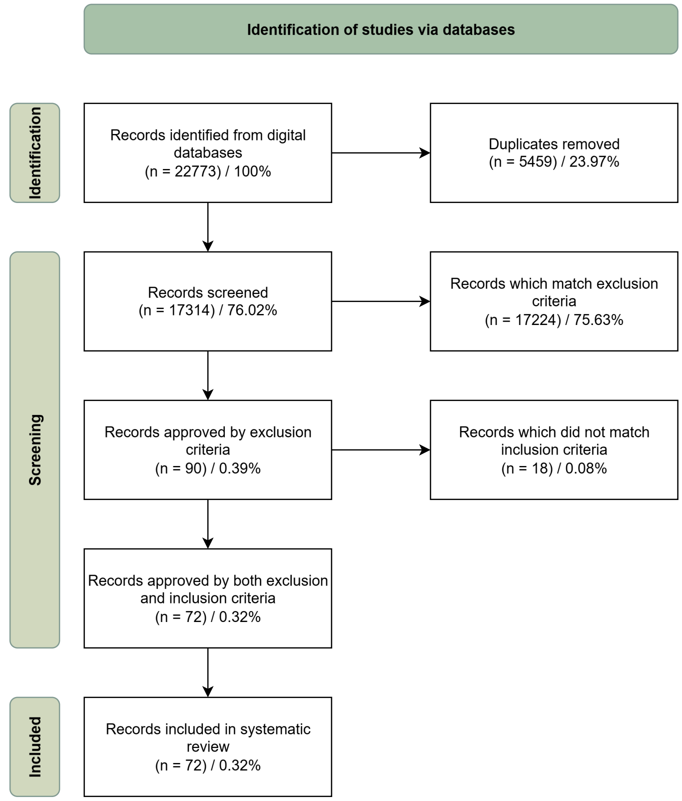

Fig. 1. Flowchart of studies selection

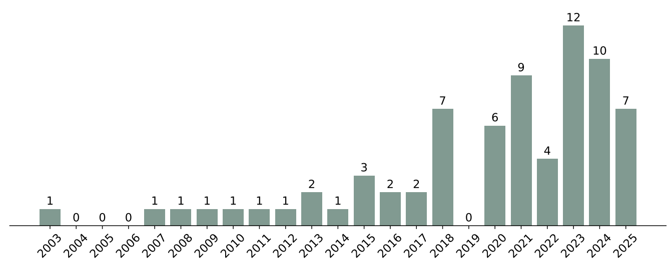

Fig. 2. Distribution of studies by year

B. Quality assessment

When it comes to quality evaluation, 18 out of the 72 included studies achieved the maximum possible quality score of 4, with 62 studies scoring at least 2. Those with lower scores were generally papers that did not present a direct application of virtual reality, focusing instead on discussions of specific aspects related to such applications. The amount of studies for each quality score can be seen in Figure 3.

The analysis of the selected studies revealed substantial variation in both assessment methods and healthcare domains. Three major themes arise from this review: the diversity found in evaluation methods, the concentration of studies in certain healthcare fields, and the overall trends and quality standards of research in this area.

The review showed that a significant proportion of studies attained high scores based on the stated quality criteria,

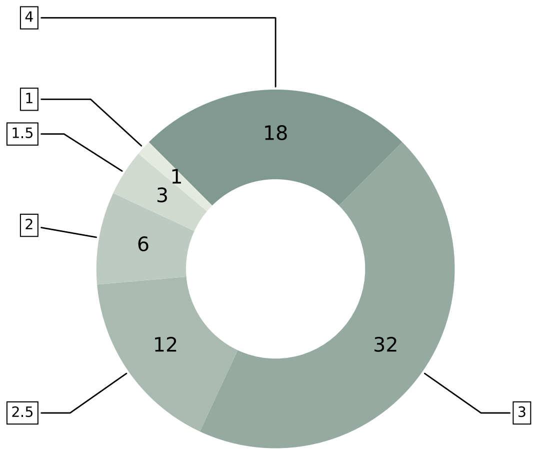

Fig. 3. Amount of studies by quality score

with a quarter receiving top marks. This result suggests that, despite the relatively recent rise in accessible VR technologies, research conducted in this space demonstrates a relatively strong level of methodological soundness and reporting rigor. However, some studies lacked adequate detail regarding their assessment protocols, which compromised their reproducibility.

Ensuring the reproducibility of scientific findings is a major pillar of ongoing research progress. Therefore, future investigations in this domain should emphasize transparent and comprehensive reporting of study procedures, especially when introducing new types of metrics, assessment tools, or AI-based analysis techniques. The development of international reporting guidelines could further improve the robustness and comparability of research in VR-based health education.

C. Learning evaluation methods

A relatively small proportion of the reviewed papers employed exclusively subjective approaches for learning evaluation. Figure 4 illustrates the distribution of papers by evaluation method category. Approximately 87.5% of the studies fell into either the ”Objective” or ”Both” category, indicating that the large majority of studies incorporated at least one objective method to assess learning.

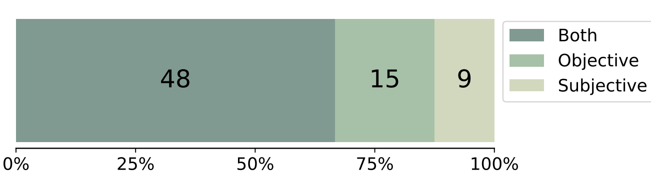

Fig. 4. Learning evaluation methods in the included articles

D. Healthcare fields

In terms of the healthcare disciplines represented, only a handful of distinct fields appeared among the selected papers, as shown in Figure 5.

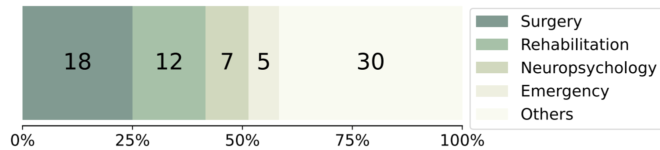

Fig. 5. Healthcare fields present in selected studies

The fields most frequently explored were surgery (with an emphasis on surgical training for diverse procedures) [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41], rehabilitation (particularly related to the recovery of stroke patients) [42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53], and neuropsychology [54, 55, 56, 57, 58, 59, 60]. Additional areas included emergency medicine, dentistry, cardiology, and anatomy.

It is worth noting that particular domains, especially surgery and rehabilitation, appear to be using VR more than others. Surgical training can take great advantage of VR’s ability to simulate intricate movements and offer repeated, risk-free practice outside of patient care.

Within rehabilitation, VR is frequently used to facilitate both motor and cognitive recovery through engaging, gamified, and adaptable tasks [42, 46]. Here, learning evaluation occasionally incorporates physiological or behavioral metrics, such as range of motion or adherence to program tasks. In both of these major domains, the capabilities of VR enable forms of skill development that are difficult to achieve with traditional training modules.

By contrast, areas such as anatomy, pediatrics, geriatrics, and ophthalmology were underrepresented, pointing to opportunities for expansion in these fields. The further application of VR in such areas could be particularly beneficial as the underlying technology continues to become more accessible and user-friendly. Nevertheless, the current limited evidence in these domains poses challenges for the creation of standardized evaluation methods. Future studies are needed to extend VR-based medical education and to adapt or replicate successful assessment models from other fields.

E. Type of approach for evaluation

It is noticeable that only a limited number of studies have adopted artificial intelligence (AI) tools for evaluation purposes, as Figure 6 illustrates. The majority of papers continued to rely on traditional statistical analysis methods. The utilization of AI-based assessment techniques has risen in studies published between 2018 and 2024 [34, 36, 61, 62].

Aside from two studies [38, 63], almost all pursued primarily quantitative assessment strategies, often paired with qualitative approaches for richer analysis.

There is an emerging and important trend regarding the use of AI in analyzing learning outcomes. From 2003 to 2017, the average number of papers using AI was essentially zero, but this number rose to 0.57 per year during 2018-2025. Although

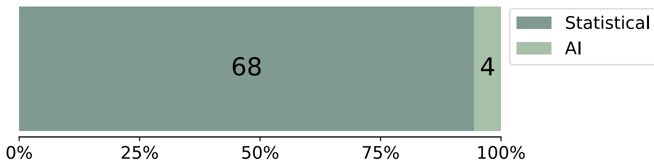

Fig. 6. Type of approach in selected studies

most studies still use more conventional statistical methods, an increasing subset is experimenting with machine learning to analyze user behaviors or adapt learning experiences to individual needs.

AI methods enable analysis of the complex, high-volume data generated during VR sessions, including hand tracking, gaze monitoring, and physiological indicators. These analyses can potentially provide highly adaptive feedback and individualized insight into each learner's process. Despite these advantages, such methods raise important questions about validity, interpretability, and transparency, especially in cases where there is no external benchmark for result validation.

Moreover, despite their promise, AI-driven evaluation approaches remain largely limited to pilot or proof-of-concept studies at this stage. For wider adoption, more collaboration will be needed between researchers, clinical professionals, and AI developers to advance tools that are both technically sound and pedagogically meaningful.

F. Type of VR technologies

Despite the unique objectives of each study, several technologies were common to multiple papers. The software and hardware technologies that appeared repeatedly are summarized in Tables III and IV.

TABLE III
SOFTWARE USED

Technology	Number of papers
Unity3D	31 (43.0%)
SteamVR	7 (9.7%)
Blender	7 (9.7%)
Autodesk Maya	5 (6.9%)
OpenGL	4 (5.5%)
OpenHaptics	3 (4.2%)
C++	3 (4.2%)

The importance of these software packages is underscored by their widespread adoption in the VR healthcare training literature. Most notably, Unity3D stands out as the core technology for more than 40% of included studies. Other commonly used applications include Blender and Autodesk Maya, which are essential for generating detailed 3D models used in VR simulations.

Similarly, Table IV points to the dominance of the Meta Oculus and HTC Vive as preferred hardware, with nearly half (44.4%) of evaluated papers employing one or both.

TABLE IV
HARDWARE USED

Technology	Number of papers
Meta Oculus	20 (27.8%)
HTC Vive	12 (16.7%)
PHANToM Haptic Devices	7 (9.72%)
Google Cardboard	3 (4.2%)
Cave Automatic Virtual Environment	3 (4.2%)

G. Evaluation types

Given the diversity of ways in which evaluation can be performed, the included studies were categorized into three primary evaluation types: Effect (assessing the impact of VR experience on user performance and task improvement), Functional (focusing on aspects such as application usability, user experience, and workflow), and Technical (examining technical characteristics such as system latency). The frequency of each evaluation approach is presented in Table V.

TABLE V
TYPES OF EVALUATION

Evaluation	Number of papers
Effect	45 (62.5%)
Functional	26 (36.1%)
Technical	1 (1.39%)

As summarized in Table V, a significant majority of studies prioritized measuring the effect of the intervention on the end user's learning, but many also assessed the functional quality or technical parameters of VR applications.

H. Collected metrics

The core of learning evaluation lies in the collection of relevant metrics, which allows experimenters to quantitatively verify whether educational objectives are being achieved. In the studies examined, data collection was primarily conducted either through a combination of manual and automatic methods (in 31 studies) or exclusively automated processes (in 20 studies). Only a minority of studies depended solely on manual data gathering.

Given the vast variety of metrics tracked across all studies, these were consolidated into broad categories as suggested by [64], and outlined in Table VI. Furthermore, Figure 7 visually depicts the distribution of metric categories by healthcare field.

TABLE VI
CATEGORY OF METRICS COLLECTED

Category	Number of papers
Time	42 (58.3%)
Error or Success	33 (45.8%)
Precision	13 (18.1%)
Movement	12 (16.7%)
Force	9 (12.5%)

As the data in Table VI indicates, the measurement of time was the most prevalent metric. This frequently included

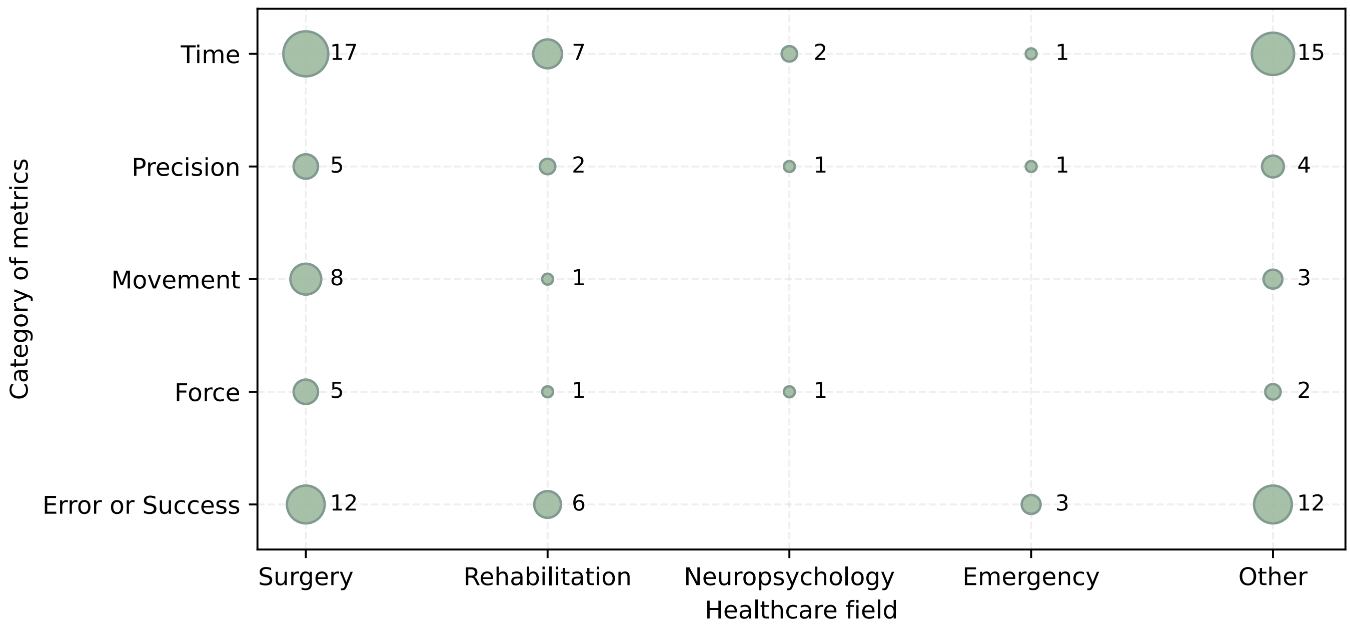

Fig. 7. Amount of papers by category of metrics and healthcare field

metrics such as the time taken to complete a task, which was evaluated in 24 studies under the Effect category. Movement was also a commonly used metric, typically being comprised of aspects such as proper trajectory adherence during a task, speed, and instance penalties for collision events. Several representative examples are listed in Table VII.

TABLE VII
EXAMPLES OF METRICS COLLECTED

Category	Examples of metrics
Time	Task completion time; Time to answer (a question); Total experiment time; Latency
Error or Success	Collision with sensitive structures; Number of errors; Error rate; Correct answers; Measurement error
Precision	Diagnosis accuracy; Calibration accuracy; Overlap rate
Movement	Acceleration; Speed; Jerk; Economy of movements; Path length
Force	Average force applied to a tissue; Torque magnitude; Pressure on fabric

One key observation from the review is the pronounced lack of standardization in the assessment of learning within VR-based healthcare training. The studies reported a broad spectrum of evaluation instruments, including written knowledge tests, various performance indicators (such as time and accuracy), participant self-reports on confidence, and even physiological or behavioral data automatically gathered by the VR application.

While this diversity highlights the diverse nature of learning (spanning cognitive, behavioral, and affective domains), it also imposes substantial barriers to comparison between studies and makes the execution of meta-analyses more difficult. Some papers emphasized learning as measured by objective data (for example, task completion time, surgical precision, error rates), whereas others coupled these with subjective feedback on participant perceptions, whether concerning realism or general satisfaction. Standardized questionnaires like the System Usability Scale (SUS) appeared in seven studies, and others, such as the Game Experience Questionnaire (GEQ) [65, 66], the Objective Structured Clinical Examination (OSCE) [67, 68], and the Lawton Brody Instrumental Activities of Daily Living Scale (IADL) [50, 69], were also employed.

This wide-ranging use of tools and metrics reveals a critical gap: the absence of unified and validated frameworks specifically adapted to immersive virtual learning environments. Traditional assessment instruments, such as pre- and post-intervention knowledge tests, are often insufficient to fully capture the amount and depth of learning achieved in VR contexts. This highlights the urgent need for validated assessment tools designed specifically to reflect the unique characteristics of learning within VR, especially as these technologies become more mainstream in health education.

I. *Experiments with users*

Several studies tested their VR applications or assessment methods not just with the eventual end users (such as students or trainees) but also with intermediate users, including lab staff or faculty instructors. The breakdown of test participant types is shown in Figure 8.

Within the studies examined, the median number of end users participating in the tests was 30, and the typical number of instructors involved as testers was 2.

J. *Limitations of the review*

This systematic review is subject to several limitations. First, the search was conducted using only five digital databases and

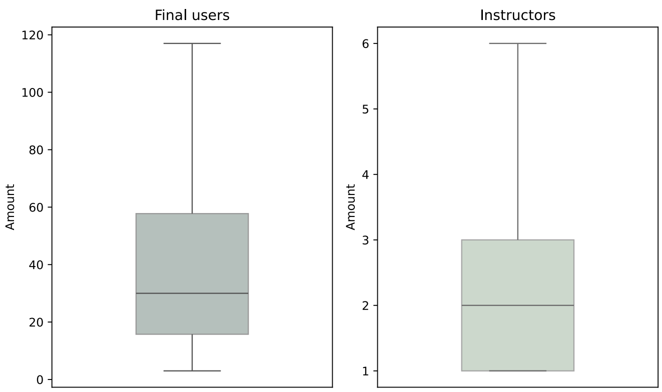

Fig. 8. Distribution of amount of users by type

applied a relatively restrictive search string, which may have inadvertently excluded some relevant publications, particularly those from educational or clinical fields not indexed in the selected databases. Second, the substantial heterogeneity in study designs and variables limited the scope for meta-analysis or quantitative synthesis. Finally, despite conscientious effort in evaluating study quality, some degree of subjectivity is inherent in this process, particularly regarding the choice and weighting of assessment questions.

K. Challenges and opportunities

From what was said in previous sections, it is possible to conclude that despite the progress observed in the development and evaluation of VR-based training in healthcare, there are still a few challenges and many research opportunities. One of the main difficulties is the creation of real-time assessment methods, which can provide immediate feedback and support adaptive and personalized learning during the training sessions, as most methods still rely on data analysis after the learning session. Furthermore, there is a lack of reliable and standardized methods capable of measuring learning outcomes with consistency across different training scenarios. Another challenge involves designing intelligent systems that can automatically adapt to the learner's performance, modifying the training environment or tasks based on ongoing assessments.

These and other limitations represent valuable research opportunities for the future. Addressing them can lead to more effective, personalized, and scalable VR-based training solutions. Some of the main opportunities include:

- Development of standardized assessment frameworks: Conceptual frameworks and/or implemented systems can be designed to include well-defined phases that are common to healthcare simulators. These frameworks should support the consistent evaluation of learning outcomes across different VR applications and contribute to better comparisons between studies.
- Exploration of underrepresented areas in healthcare: Several healthcare fields remain underexplored in the context of VR-based education. These domains often provide access to rich data sources, including clinical records, patient-reported outcomes, and medical imaging, which could be used to support the design and evaluation of three-dimensional training experiences.
- Real-time assessment methods: Many of the current evaluation strategies are applied only after the training session has ended. Although this approach allows for detailed analysis, it may miss important insights that could be captured during the learning process. Real-time assessment methods can provide immediate feedback to learners, allowing them to adjust their behavior or understanding during the task. These methods can also support adaptive learning systems that respond dynamically to learner progress.
- Reliable and robust evaluation techniques: When learning-based methods are involved in the assessment process, it is important to ensure that the models are trained on large, diverse datasets that reflect the real-world variability of healthcare learners and scenarios. This is essential to improve the reliability, fairness, and transparency of AI-driven or other learning-driven evaluations, and to avoid biases that could negatively affect learners or educational outcomes.
- Ethical and practical implications of AI-based assessments: As AI becomes more integrated into educational tools, it is important to address concerns related to learner privacy, data protection, algorithmic fairness, accountability and the explainability of automated decisions. Future studies should examine how these factors affect the acceptance and effectiveness of AI-based assessment tools in clinical training.

Taking all into account, it is evident that a lot remains to be explored and developed in this field. By working on these opportunities, researchers and developers may overcome the current challenges in the field of VR-based healthcare education, making it more effective, inclusive, and ethical.

V. Conclusion

A major finding is the marked diversity of assessment approaches. Studies utilized a combination of objective performance indicators (like task completion time, precision, and error rate), subjective measures (including self-report surveys on satisfaction and realism), as well as newer forms of automatic data collection from user interaction or physiological sensors. The catalog of evaluation techniques discussed in Table VII directly addresses the research question, illustrating both preferred metrics for assessing learning in healthcare VR studies and their broader categorical groupings. Although this amount reflects the inherent complexity of learning in VR settings, it also represents a significant obstacle: the current lack of common, validated evaluation frameworks adapted specifically for immersive educational environments. In the absence of such standards, it becomes difficult to directly compare outcomes across studies or to perform rigorous meta-analyses, therefore making the overall progress of evidence-based practice in this area more difficult.

Another important observation is the predominance of VR applications in specific branches of healthcare, most notably surgery and rehabilitation. These fields benefit substantially from VR's capacity to deliver high-fidelity, interactive simulations that allow repeated skills practice without exposing patients to any risk. However, other potentially promising areas, such as anatomy, mental health, home care, and professional communication, are far less represented. Expanding VR's use into these domains may enrich the scope and impact of immersive learning but will also require the development of both technological solutions and field-specific evaluation strategies.

When considering technological trends, the increasing integration of AI and machine learning represents an emerging direction for individualized and large-scale assessment. While present applications remain mostly exploratory, AI's capacity to analyze complex, high-volume behavioral data from VR sessions opens new opportunities for monitoring, feedback, and adaptation. Still, as the use of AI becomes more widespread, further consideration must be given to ethical matters, including data privacy, security, fairness and accountability.

Lastly, quality assessment of the reviewed studies indicates a generally positive trend: most studies demonstrated high standards for methodological rigor, reproducibility, and the potential relevance of findings. Nevertheless, opportunities remain for improvements in consistent reporting, sharing of evaluation protocols, and strengthening the reproducibility of results.

References

[1] Bhone Myint Kyaw et al. "Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration". In: *Journal of medical Internet research* 21.1 (2019), e12959.

[2] Irene Alice Chicchi Giglioli et al. "Augmented reality: a brand new challenge for the assessment and treatment of psychological disorders". In: *Computational and mathematical methods in medicine* 2015.1 (2015), p. 862942.

[3] David A Cook et al. "Technology-enhanced simulation for health professions education: a systematic review and meta-analysis". In: *Jama* 306.9 (2011), pp. 978–988.

[4] Jingjie Zhao et al. "The effectiveness of virtual reality-based technology on anatomy teaching: a meta-analysis of randomized controlled studies". In: *BMC medical education* 20 (2020), pp. 1–10.

[5] Feng-Qin Chen et al. "Effectiveness of virtual reality in nursing education: meta-analysis". In: *Journal of medical Internet research* 22.9 (2020), e18290.

[6] Henna Mäkinen et al. "User experiences of virtual reality technologies for healthcare in learning: an integrative review". In: *Behaviour & Information Technology* 41.1 (2022), pp. 1–17. DOI: 10.1080/0144929X.2020.1788162. eprint: https://doi.org/10.1080/0144929X.2020.1788162. URL: https://doi.org/10.1080/0144929X.2020.1788162.

[7] Hyunkyung Sung et al. "Effectiveness of Virtual Reality in Healthcare Education: Systematic Review and Meta-Analysis". In: *Sustainability* 16.19 (2024). ISSN: 2071-1050. DOI: 10.3390/su16198520. URL: https://www.mdpi.com/2071-1050/16/19/8520.

[8] Fabiana Frata Furlan Peres et al. "Methods for Evaluating Immersive 3D Virtual Environments: a Systematic Literature Review". In: *Proceedings of the 26th Symposium on Virtual and Augmented Reality*. SVR '24. Manaus, Brazil: Association for Computing Machinery, 2024, 140–151. ISBN: 9798400709791. DOI: 10.1145/3691573.3691595. URL: https://doi.org/10.1145/3691573.3691595.

[9] Sahra K G Silva et al. "A framework for evaluating depth perception in stereoscopic virtual environments". In: *Proceedings of the 24th Symposium on Virtual and Augmented Reality*. SVR '22. Natal, RN, Brazil: Association for Computing Machinery, 2024, 105–114. ISBN: 9798400700026. DOI: 10.1145/3604479.3604529. URL: https://doi.org/10.1145/3604479.3604529.

[10] Yuri Fernando Tiecher Sefrin, Claudio Roberto Marquetto Mauricio, and Fabiana Frata Furlan Peres. "Usability Evaluation of a 3D Digital-TwinIndirect usability evaluation of SOMA – System Oriented to Monitoring of Engineering Assets". In: *Proceedings of the 23rd Symposium on Virtual and Augmented Reality*. SVR '21. Virtual Event, Brazil: Association for Computing Machinery, 2022, 157–161. ISBN: 9781450395526. DOI: 10.1145/3488162.3488220. URL: https://doi.org/10.1145/3488162.3488220.

[11] Filipe A Fernandes et al. "Evaluating User Experience of a Software Engineering Education Virtual Environment". In: *Proceedings of the 24th Symposium on Virtual and Augmented Reality*. SVR '22. Natal, RN, Brazil: Association for Computing Machinery, 2024, 137–141. ISBN: 9798400700026. DOI: 10.1145/3604479.3604516. URL: https://doi.org/10.1145/3604479.3604516.

[12] Esther Z Barsom, Maurits Graafland, and Marlies P Schijven. "Systematic review on the effectiveness of augmented reality applications in medical training". In: *Surgical endoscopy* 30 (2016), pp. 4174–4183.

[13] Jaziar Radianti et al. "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda". In: *Computers & education* 147 (2020), p. 103778.

[14] Jeremy N Bailenson et al. "Equilibrium theory revisited: Mutual gaze and personal space in virtual environments". In: *Presence: Teleoperators & Virtual Environments* 10.6 (2001), pp. 583–598.

[15] Mel Slater, Sylvia Wilbur, et al. "A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments". In:

Presence: Teleoperators and virtual environments 6.6 (1997), pp. 603–616.
[16] Bob G Witmer and Michael J Singer. “Measuring presence in virtual environments: A presence questionnaire”. In: *Presence* 7.3 (1998), pp. 225–240.
[17] Tainã Ribeiro de Oliveira et al. “Virtual reality solutions employing artificial intelligence methods: A systematic literature review”. In: *ACM Computing Surveys* 55.10 (2023), pp. 1–29.
[18] Christopher J Dede, Jeffrey Jacobson, and John Richards. “Introduction: Virtual, augmented, and mixed realities in education”. In: *Virtual, augmented, and mixed realities in education*. Springer, 2017, pp. 1–16.
[19] Ronei M Moraes and Liliane S Machado. “Fuzzy continuous evaluation in training systems based on virtual reality”. In: *Proc. of 2009 IFSA World Congress, Lisboa*. 2009, pp. 102–107.
[20] Andrew Gelman et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
[21] Zahira Merchant et al. “Effectiveness of virtual reality-based instruction on students’ learning outcomes in K-12 and higher education: A meta-analysis”. In: *Computers & education* 70 (2014), pp. 29–40.
[22] Jessica Isbely Montana et al. “Neurorehabilitation of spatial memory using virtual environments: a systematic review”. In: *Journal of clinical medicine* 8.10 (2019), p. 1516.
[23] Doug A Bowman and Ryan P McMahan. “Virtual reality: how much immersion is enough?” In: *Computer* 40.7 (2007), pp. 36–43.
[24] Sergio Teodoro-Vite et al. “A high-fidelity hybrid virtual reality simulator of aneurysm clipping repair with brain sylvian fissure exploration for vascular neurosurgery training”. In: *Simulation in Healthcare* 16.4 (2021), pp. 285–294.
[25] José Ignacio Guzmán, Mauricio Herrera, and Camilo Rodríguez Beltrán. “A New Highly Portable Simulator (SECMA) Based on Virtual Reality for Teaching Essential Skills in Minimally Invasive Surgeries.” In: *International Journal of Online & Biomedical Engineering* 19.13 (2023).
[26] Jun Liu et al. “Cloud computing-enabled IIOT system for neurosurgical simulation using augmented reality data access”. In: *Digital Communications and Networks* 9.2 (2023), pp. 347–357.
[27] Anas Amin Preukschas et al. “Comparing a virtual reality head-mounted display to on-screen three-dimensional visualization and two-dimensional computed tomography data for training in decision making in hepatic surgery: A randomized controlled study”. In: *Surgical Endoscopy* 38.5 (2024), pp. 2483–2496.
[28] Reza Karimzadeh et al. “Design and implementation of brain surgery bipolar electrocautery simulator using haptic technology”. In: *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 47.3 (2023), pp. 859–869.
[29] H Çakmak et al. “Haptic ventriculostomy simulation in a grid environment”. In: *Computer animation and virtual worlds* 20.1 (2009), pp. 25–38.
[30] Sandrine De Ribaupierre et al. “Healthcare training enhancement through virtual reality and serious games”. In: *Virtual, augmented reality and serious games for healthcare 1* (2014), pp. 9–27.
[31] Seunghan Lee, Amar Sadanand Shetty, and Lora A Cavuoto. “Modeling of learning processes using continuous-time markov chain for virtual-reality-based surgical training in laparoscopic surgery”. In: *IEEE transactions on learning technologies* 17 (2023), pp. 462–473.
[32] Dariusz Łaski et al. “New comprehensive surgical curriculum of pre-graduate surgical education”. In: *Videosurgery and Other Miniinvasive Techniques* 8.3 (2013), pp. 200–210.
[33] Louise Moody et al. “Objective metrics for the evaluation of simple surgical skills in real and virtual domains”. In: *Presence: teleoperators & virtual environments* 12.2 (2003), pp. 207–221.
[34] Fei Li et al. “Personalized assessment and training of neurosurgical skills in virtual reality: An interpretable machine learning approach”. In: *Virtual Reality & Intelligent Hardware* 6.1 (2024), pp. 17–29.
[35] Costas S Tzafestas et al. “Pilot evaluation study of a virtual paracentesis simulator for skill training and assessment: The beneficial effect of haptic display”. In: *Presence: Teleoperators and Virtual Environments* 17.2 (2008), pp. 212–229.
[36] Paulo VF Paiva et al. “SimCEC: a collaborative VR-based simulator for surgical teamwork education”. In: *Computers in Entertainment (CIE)* 16.2 (2018), pp. 1–26.
[37] Junzhen Du et al. “Toward Immersive and Interactive Surgical Training Using Extended Reality Simulator for IoMT”. In: *International Journal of Human–Computer Interaction* 41.8 (2025), pp. 4833–4850.
[38] Vuthea Chheang et al. “Toward interprofessional team training for surgeons and anesthesiologists using virtual reality”. In: *International journal of computer assisted radiology and surgery* 15 (2020), pp. 2109–2118.
[39] Jun J Pan et al. “Virtual reality training and assessment in laparoscopic rectum surgery”. In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 11.2 (2015), pp. 194–209.
[40] Amir Amini et al. “Virtual Reality vs Phantom Model: Benefits and Drawbacks of Simulation Training in Neurosurgery”. In: *Operative Neurosurgery* 27.5 (2024), pp. 618–631.
[41] Vasileios Lahanas et al. “Virtual reality-based assessment of basic laparoscopic skills using the Leap Motion controller”. In: *Surgical endoscopy* 31 (2017), pp. 5012–5023.
[42] Aline Menin, Rafael Torchelsen, and Luciana Nedel. “An analysis of VR technology used in immersive simu-

lations with a serious game perspective". In: *IEEE computer graphics and applications* 38.2 (2018), pp. 57–73.

[43] Simone Gazzellini et al. "Clinical study and Health Technology Assessment (HTA) of a Robot-Assisted Gait Training on children with neurological disorders: A quasi-experimental study". In: *J Health Soc Sci* 8.1 (2023), pp. 76–92.

[44] Won-Jun Jeong et al. "Establishment of production standards for web-based metaverse content: focusing on accessibility and HCI". In: *Journal of Web Engineering* 21.8 (2022), pp. 2231–2256.

[45] Ashok Kumar Patil et al. "Fusion of multiple lidars and inertial sensors for the real-time pose tracking of human motion". In: *Sensors* 20.18 (2020), p. 5342.

[46] Roy Francis Navea et al. "Gamified Shoulder Rehabilitation for Mild Stroke Patients Using Virtual Reality." In: *International Journal of Technology* 16.1 (2025).

[47] Patricia Capsi-Morales et al. "Impact of Visual Feedback Configurations in a Task-oriented Immersive Virtual Reality Mirror Therapy". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024).

[48] Hao Zhou et al. "Learning on the rings: Self-supervised 3D finger motion tracking using wearable sensors". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.2 (2022), pp. 1–31.

[49] Maria Matsangidou et al. "Participatory design and evaluation of virtual reality physical rehabilitation for people living with dementia". In: *Virtual Reality* 27.1 (2023), pp. 421–438.

[50] Murat Akinci et al. "The Effects of Combined Virtual Reality Exercises and Robot Assisted Gait Training on Cognitive Functions, Daily Living Activities, and Quality of Life in High Functioning Individuals With Subacute Stroke". In: *Perceptual and Motor Skills* 131.3 (2024), pp. 756–769.

[51] Van-Hanh Nguyen, Frédéric Merienne, and Jean-Luc Martinez. "Training based on real-time motion evaluation for functional rehabilitation in virtual environment". In: *International Journal of Image and Graphics* 10.02 (2010), pp. 235–250.

[52] Daria Tsoupikova et al. "Virtual immersion for post-stroke hand rehabilitation therapy". In: *Annals of biomedical engineering* 43 (2015), pp. 467–477.

[53] Mindy F Levin et al. "Virtual reality versus conventional treatment of reaching ability in chronic stroke: clinical feasibility study". In: *Neurology and therapy* 1 (2012), pp. 1–15.

[54] Christoph Rockstroh, Johannes Blum, and Anja S Göritz. "A mobile VR-based respiratory biofeedback game to foster diaphragmatic breathing". In: *Virtual Reality* 25.2 (2021), pp. 539–552.

[55] Stavros Skouras et al. "Earliest amyloid and tau deposition modulate the influence of limbic networks during closed-loop hippocampal downregulation". In: *Brain* 143.3 (2020), pp. 976–992.

[56] Horace HS Ip et al. "Enhance emotional and social adaptation skills for children with autism spectrum disorder: A virtual reality enabled approach". In: *Computers & Education* 117 (2018), pp. 1–15.

[57] Rashmita Chatterjee and Zahra Moussavi. "Evaluation of a cognition-sensitive spatial virtual reality game for Alzheimer's disease". In: *Medical & Biological Engineering & Computing* (2024), pp. 1–11.

[58] Sun Kyung Kim et al. "Expanding virtual reality simulation with reflective learning to improve mental health nursing skills of undergraduate nursing students". In: *Education and Information Technologies* (2024), pp. 1–25.

[59] Yingjie Li et al. "Haptic-based virtual environment design and modeling of motor skill assessment for brain injury patientsrehabilitation". In: *Computer-Aided Design and Applications* 8.2 (2011), pp. 149–162.

[60] Anna Junga et al. "Teach the Unteachable with a Virtual Reality (VR) Brain Death Scenario–800 Students and 3 Years of Experience". In: *Perspectives on Medical Education* 14.1 (2025), p. 44.

[61] Nicholas Ho et al. "Virtual reality training for assembly of hybrid medical devices". In: *Multimedia Tools and Applications* 77 (2018), pp. 30651–30682.

[62] Xiaodan Yan et al. "Privacy preserving for AI-based 3D human pose recovery and retargeting". In: *ISA transactions* 141 (2023), pp. 132–142.

[63] Jordi Mill et al. "Domain expert evaluation of advanced visual computing solutions and 3D printing for the planning of the left atrial appendage occluder interventions". In: *International journal of bioprinting* 9.1 (2022), p. 640.

[64] Lucas H. Sallaberry, Romero Tori, and Fátima L. S. Nunes. "Automatic Performance Assessment in Three-dimensional Interactive Haptic Medical Simulators: A Systematic Review". In: *ACM Comput. Surv.* 55.7 (Dec. 2022). ISSN: 0360-0300. DOI: 10.1145/3539222. URL: https://doi.org/10.1145/3539222.

[65] Won-Jun Jeong et al. "Establishment of production standards for web-based metaverse content: focusing on accessibility and HCI". In: *Journal of Web Engineering* 21.8 (2022), pp. 2231–2256.

[66] W.A. IJsselsteijn, Y.A.W. de Kort, and K. Poels. *The Game Experience Questionnaire*. English. Technische Universiteit Eindhoven, 2013.

[67] Marvin Mergen et al. "Feasibility study on virtual reality-based training for skin cancer screening: Bridging the gap in dermatological education". In: *Education and Information Technologies* 30.4 (2025), pp. 5251–5282.

[68] Bouthina K Greiw. "Objective Structured Clinical Examination (OSCE)". In: *Misurata Medical Sciences Journal* (2015), p. 147.

[69] Jessica Fish. "Lawton-Brody instrumental activities of daily living scale". In: *Encyclopedia of clinical neuropsychology*. Springer, 2011, pp. 1438–1439.