Evaluation Methods Applied to Virtual Reality Educational Applications: A Systematic Review

Daniel Bueno Domingueti[1(✉)], Diego Roberto Colombo Dias[1], Marcelo de Paiva Guimarães[2], and Dárlinton Barbosa Feres Carvalho[1]

[1] Universidade Federal de São João del-Rei, São João del-Rei, Brazil
danieldomingueti@aluno.ufsj.edu.br,
{diegodias,darlinton}@ufsj.edu.br
[2] Universidade Federal de São Paulo/UNIFACCAMP, Campo Limpo Paulista, Brazil
marcelo.paiva@unifesp.br

Abstract. Evaluation is an important activity in software development since it allows detecting issues like design flaws under different analysis perspectives. However, conducting a satisfactory assessment is challenging, especially considering new interactive technologies such as virtual reality (VR). This paper presents a systematic review revealing the most used evaluation methods in this context to deepen the understanding of the evaluation of educational applications employing VR. Considering a search in the ACM database with a filter for the last six years, we selected 1351 initial studies and 81 for analysis. The results show that the tests of usability, flow, and technology acceptance model are the most common, as well as the methods used in many articles present weaknesses due to the lack of an adequate theoretical foundation. Accordingly, we strengthen the relevance of using well-defined support theories in all stages of an evaluation.

Keywords: Evaluation · Virtual Reality · Systematic review

1 Introduction

Evaluation is a crucial stage in software development, being considered in various methodologies and development processes [31]. At this stage, it is possible to assess the product value and quality, and in the case of research projects, it enables us to check the hypothesis and draw conclusions. An evaluation can be executed based on different perspectives, each having its principles and methods.

From another perspective, the interplay between users and software applications has been studied by the Human-Computer Interaction (HCI). This field considers a set of elements for analysis in a software evaluation, but mainly regarding user experience and design alternatives.

Each evaluation perspective and the different kinds of software interactions bring a variety of methods and techniques available to perform the evaluation stage. There are numerous approaches in the literature for such purposes.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12958, pp. 641–657, 2021.
https://doi.org/10.1007/978-3-030-87016-4_46

Nevertheless, multiple perspectives are relevant when analyzing a software product to get a better quality assessment. However, building an evaluation instrument is not a trivial task, requiring many steps and deliberations to be performed. For instance, related to actions to conduct an evaluation, the *DECIDE Framework* establishes a formal structure to create and execute a user evaluation [29] in six steps.

Defining an evaluation instrument can be even more challenging when dealing with new (or unique) interaction styles such as mixed reality systems—a concept that joins Virtual Reality (VR) with Augmented Reality (AR). The challenge concerning those peculiar evaluations includes the greater need for specialized resources as well as concerning the definition of the task for analysis—mainly due to the unusual interaction style adopted by the advanced applications [2].

Notwithstanding, the new interaction styles provided by VR brought a set of significant innovations to many research areas and commercial applications. Notably, VR is scoring remarkable results in the context of educational software. For example, it can help with medical training and education [25], and even providing new experiences on education [33].

Accordingly, this paper presents research with the objective of providing a better understanding of how educational software applications based on VR have been evaluated through a systematic review. This study seeks to identify and characterize the most popular research approaches and evaluation methods. Thus, we expect to promote a better comprehension of how evaluation methods have been used in research considering VR applications aimed at education. Consequently, we hope it nurtures better practices in dealing with the evaluation of new technologies.

2 Background

The experimental evaluation stage plays a vital role in software development and scientific research. Experiments are conducted to reveal the influences of intervention in such a way as to allow varying input parameters and learn possible causes, effects, and causal relations [28]. This stage has evolved over time based on scientific thinking: a post-positivist philosophy that involves the exam and interpretation of pieces of evidence into a theoretical structure molding all the scientific research stages ([16] *apud* [17]).

Four main elements define the experimental evaluation, as depicted in Fig. 1: support theories, method, data collection tools, and results. Determined by the practiced method, the research type (or approach) fixes the range of possible data collection tools. Moreover, data collection tools dictate how the data will be analyzed. According to scientific thinking and method, the whole process should be anchored at well-defined support theories.

A method defines a structured process and guides researchers through the whole evaluation performance. That is, it determines how the evaluation (and the research) will be conducted, which steps will be performed, and the order of their execution [12]. The chosen method is related to the study objective

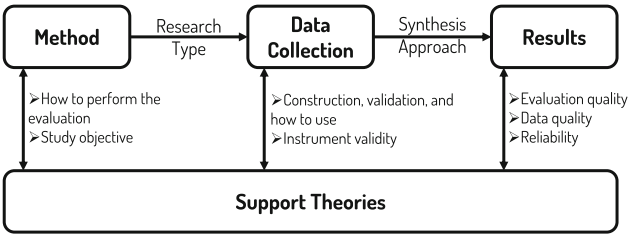

Fig. 1. The four main elements used to characterize an experimental evaluation and their relationships.

and must be based on a well-defined academic grounding. Thus, as long as the selected method counts on reliable supporting theories, the experiment results deliver scientific value. That is, the extent to which the results are valuable and contributing to the body of knowledge, especially from the current experiment scenario to other contexts [8] *apud* [1].

Therefore, the chosen method, along with the study objectives, establishes an approach qualifying the experiment. The paradigm adopted by the researcher (i.e., the primary research approach) refers to how the researcher perceives and interprets the world. The research approach may be qualified as quantitative, qualitative, mixed, or, as defined in this study, combined. The following paradigm guides the researcher in defining or building the data collection tools based on the data types to be collected further.

The mixed approach uses both classical approaches (i.e., quantitative and qualitative) in order to provide a better comprehension of the evaluated phenomena through a complementary (or even a different) point of view [10]. In our study, however, we use the concept of combined approach: the research uses both classical approaches, although it is not necessarily based on the mixed method theory, as described in the literature.

The use of well-defined support theories is also desirable to set the data collection process. The gold standard is to validate the data collection through statistical methods besides establishing specific instructions about its usage in experiments. Performing the validation through statistics provides a better internal quality of the test - the condition in which the variables can be related to each other, not representing measurement errors [28]. In the case of more than one possible way of analysis, each one must also be validated [8].

With support theories ensuring internal and external qualities, that is, the validity of methods, tools, and analysis view, a greater sense of scientific rigor backs the experiment, besides the whole research. Thereby, the study conclusions are not merely assigned to the hypothesis and conceptions created by the researchers but based on substantial evidence that answer - positively or negatively - the well-defined assumptions supported by the backing theories. Thus, it is also possible to ensure the experiment and research reliability, allowing its replication and confident use of the obtained results in a similar context [28].

3 Systematic Review Process

This section describes the process employed in this study to conduct a systematic review to provide a better understanding of how educational software applications based on VR have been evaluated. As a guideline for this study, we considered the method proposed by [15], which outlines the required steps and ideas, among a set of elements and goals, to conduct a systematic review in computer science studies.

3.1 Research Questions

Five research questions (RQ) were set and considered in this study based on the main elements used to characterize an experimental evaluation. It is worth mentioning, the presence of support theories is also considered, but as a feature of methods and tools.

- **RQ1**: What are the most popular methods used to evaluate educational software based on VR?
- **RQ2**: How the methods that had been used in researches were conceived?
- **RQ3**: What are the most popular data collection tools that have been used?
- **RQ4**: How the data collection tools were conceived?
- **RQ5**: What is the most common research approach which has been used by authors?

3.2 Search Strategy

The search string was constructed incrementally, combining search terms and verifying their results. Search terms from three research domains were taken into consideration (i.e., VR, evaluation, and education). These terms were selected according to know keywords presented in the literature and related to the explored domains. In particular, the evaluation domain keywords had been extracted and adapted from [7] article, since it was needed to refine and limit the results. Finally, the search string used in this article was built using *AND* and *OR* operations, as defined in the following box.

((virtual AND (reality OR environment)) OR (immersive AND environment)) **AND** (test OR validation OR analysis OR (user AND (evaluation OR study OR studies)) OR experiment OR (pilot AND study) OR participant OR subject) **AND** (education OR teaching OR learning OR schooling)

This review explored studies recently published in a six-year window, that is, articles published in January 2014 and June 2020.

Initially, a variety of portals and digital libraries have been considered in the paper retrieval process regarding the proposed search string, such as IEEE[1], ACM Digital Library[2], Scopus[3] and Web of Science[4]. However, due to the high number of results, only the ACM Digital Library portal was used to set papers for analysis.

3.3 Selection Criteria

The whole set of articles retrieved through the search string has been considered for analysis. Going forward in the revision processes, the following inclusion criteria were considered:

- Studies that have developed a VR application considering education as the main objective; and,
- Studies that have been peer-reviewed.

Studies that fill at least one of the following (exclusion) criteria have been excluded from the analysis:

- Studies related to education, but that has not developed a VR application;
- Studies that developed a VR application, which is not related to education;
- Studies that has an evaluation, which is not related to education nor VR;
- Gray literature, for example, resumed articles, posters, thesis, and others.

For each analyzed article, the selection criteria were first applied considering its abstract. Later, if it still fulfills the selection criteria, the whole paper was thoroughly analyzed.

3.4 Data Extraction

We created a report to summarize the studies and their metadata to answer our research questions. This report contains topics about the article's metadata (e.g., title, publishing date, and others) and questions related to the research questions, such as the presence of an evaluation or assessment stage, how it was done, among other relevant issues. The report was built with the online tool Google Forms[5] to ease the data collection and management. The form also contains a field to enumerate the possible research results, according to the authors, which will also be analyzed further. The extraction data was executed by one researcher, and the data was stored in a digital spreadsheet.

[1] https://ieeexplore.ieee.org/Xplore/home.jsp.
[2] https://dl.acm.org/.
[3] https://www.scopus.com/search/form.uri?display=basic.
[4] https://apps.webofknowledge.com/.
[5] https://www.google.com/forms/about/.

3.5 Data Synthesis

The studies included in the analysis through the acceptance criteria were then summarized and analyzed. The data collected on the previous step is synthesized with a detailed report based on statistics [15]. Then, we summarized the data according to the fields that have been defined in the report: publish year, type of event, the number of studies that have an evaluation stage, and others. Regarding evaluation, we defined a set of categories relating to the base elements described by Fig. 1 in Sect. 2, being how the method was built, how the data collection tool was built, and the approach used by the authors.

4 Results

We found 1351 articles in the ACM repository. We excluded 23 because of the lack of access. From 1328 left, we excluded 1246 by the exclusion criteria. The amount of studies accepted and excluded in each stage is depicted in Fig. 2. As shown in Fig. 2, 81 articles were accepted in this revision, being six were published in magazines or journals, and 75 in conference papers. Notably, there has been a rise in the number of publications through the years (Table 1). The selected articles are listed in Table 4.

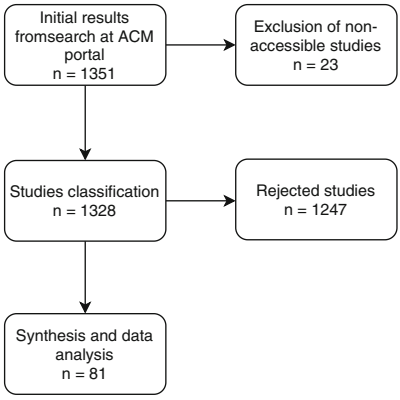

Fig. 2. Articles accepted and excluded in each stage of the systematic review

The accepted studies were classified according to the employed method, research approach, and data collection used. The classification below is applied for the method and data collection instrument of each study that includes an evaluation stage:

Table 1. Number of accepted papers grouped by year of publication

2014	2015	2016	2017	2018	2019	2020
1	3	8	13	19	28	9

- Validated: the study used a method or a data collection instrument, which is found in the literature; the study used all the rules and recommendations regarding the original authors related to the use and experimental evaluation.
- Adapted: the study used a method or data collection instrument adapted from some paper or validated research. The study may be or not be validated.
- Created: the study used a method created by the authors. As adapted studies, it may be or not be validated.

We found 61 studies that performed an evaluation stage. Only three studies have applied validated methods, 38 choose have adapted other methods (or they have some similarities with the classical ones), and 18 created their way. The validated method employed was the Quasi-Experiment method and a Mixed-method approach. However, most of the adapted evaluations were based on HCI principles. In particular instances, it is worth mentioning several studies using statistical hypothesis testing to validate the results. One used an approach based on design methodology—called interactive design by the authors—and the other article used a mixed-method approach.

Regarding the methods based on HCI principles, it is relevant to point out that the analysis was performed considering the basic experiment structure defined by the Decide Framework and the theory presented in Sect. 2. The classification was listed in Table 2.

Table 2. Articles classified by the method type

Studies	Method type	Base method
1, 3, 6, 7, 9, 11, 12, 18, 19, 21, 22, 24, 25, 35, 36, 39 ,41 ,43	Created	None
3, 5, 17, 20, 23, 28, 30, 34, 44, 46, 47, 48, 50, 52, 54,55, 57, 60, 62, 63, 65, 66, 68, 69, 72, 74, 75, 78, 81	Adapted	HCI principles
14, 31, 56, 59, 64, 71, 73, 80	Adapted	HCI principles and hypothesis test
29	Adapted	Iterative design
8, 33	Validated	Quasi-experiment
49	Validated	Mixed-method design

Relating to the 61 articles which performed an evaluation stage, we classified the approach used by the authors. The quantitative approach is the most

common, with 38 studies, while only nine articles opted for a qualitative approach. The combined method—mixing both qualitative and quantitative—has been employed by 13 articles. From all of the 60 studies that performed an evaluation stage, only two have not involved users in the experiment: study 7 has performed an experiential analysis from its authors regarding the technologies involved; and study 39 has modified the 3D environment with algorithms, describing the effects of that intervention.

Accordingly to the Data Synthesis process (Sect. 3), the evaluations were classified into two categories: simple (i.e., use only one data collection instrument); and complex (i.e., use more than one data collection instrument). Regarding these two categories, 28 articles belong to the simple type, and the rest 33 belong to the other group, the complex evaluations.

In the group of studies with simple evaluations, 22 created their data collection tool, three opted for an adapted tool, and only three used a validated tool. Only one adapted tool (43) and one created tool (8) were validated, as seen in Table 3.

Table 3. Studies grouped by tool's type and the number of validated tools considering simple evaluation studies

Tool's type	Studies	Amount of validated tools
Validated	12, 33, 78	3
Adapted	24, 28, 44	1
Created	4, 5, 6, 7, 8, 11, 17, 20, 22, 23, 35, 39, 43, 46, 55, 63, 65, 68, 71, 75	2

Studies that used validated or adapted tools have employed the following data collection tools: User Experience Questionnaire [23], Technology Acceptance Model [6], and a questionnaire extracted from a previous article [24]. Related to the created instruments, the authors used questionnaires to measure usability and user learning after the intervention, data provided for sensors (such as EEG), and statistical analysis along with their form to gauge the answer quality.

Studies performing complex evaluations employ two to five different data collection instruments. Thus, each study used a combination of validated, created, and adapted instruments, though most studies created at least one, and 11 articles changed an existing one(Fig. 3). From the adapted tools, study 18 validated a tool with the statistical process, and both studies 59 and 60 validated a tool with experts. Only one study validated a created tool (study 21).

Table 4. Primary studies selection list

ID	Title
1	Teaching Case: Applying Gamification Techniques and Virtual Reality for Learning Building Engineering 3D Arts
2	A student project experience: a virtual campus tour
3	Development and Evaluation of a Virtual Reality Driving Simulator
4	Immersive virtual reality to enhance the spatial awareness of students
5	Ogma. A Virtual Reality Language Acquisition System
6	A High-fidelity Virtual Training System for Myoelectric Prostheses Using an Immersive HMD
7	Immersive visualization anatomical environment using virtual reality devices
8	How Does Web-based Virtual Reality Affect Learning: Evidences from a Quasi-experiment
9	Passive haptics based MR system for geography teaching
10	VR biology, an interdisciplinary and international student project towards an inquiry-based pedagogy
11	Design insights into embedding virtual reality content into life skills training for people with intellectual disability
12	User perception on 3D stereoscopic cultural heritage ancient collection
13	A Virtual Experience: Benefits of Extracurricular Group Projects
14	A study on improving performance in gesture training through visual guidance based on learners' errors
15	360 Vision Applications for Medical Training
16	An Implementation of VR Chemistry Experiment System
17	Teaching-Learning Process through VR Applied to Automotive Engineering
18	Getting up your nose: a virtual reality education tool for nasal cavity anatomy
19	Behavioral Intention of Using Virtual Reality in Learning
20	Design and evaluation of multiple role-playing in a virtual film set
21	Train in virtual court: Basketball tactic training via virtual reality
22	Enhancing learning through virtual reality and neurofeedback
23	Teaching Language and Culture with a Virtual aReality Game
24	A VR Simulator for Emergency Management in Endodontic Surgery
25	Experimenting with PlayVR, a virtual reality experience for the world of theater
26	Simulator of the Glycolytic Pathway in a Virtual Reality Environment as a Learning Resource for Teaching Cellular Respiration
27	Holobody galleries: blending augmented and virtual reality displays of human anatomy
28	Exploring the potential of a 360 video application for foreign language learning
29	LearnDNA: An interactive VR application for learning DNA structure
30	Designing a virtual environment for teacher training: Enhancing presence and empathy
31	Study of self-avatar's influence on motor skills training in immersive virtual environments
32	Incorporating Virtual Reality in an Embodied Interaction Course
33	Integration of Speech Recognition into Virtual Classroom to Enhance Learning Achievement
34	A Movement Analysis System based on Immersive Virtual Reality and Wearable Technology for Sport Training
35	Drawing in a Virtual 3D Space - Introducing VR Drawing in Elementary School Art Education
36	Innovation in Urban Design Education
37	Personal Basketball Coach: Tactic Training Through Wireless Virtual Reality
38	Geometrical Hong Kong: An Immersive and Interactive Virtual Reality Tour
39	Visualizing the tape of life: exploring evolutionary history with virtual reality
40	Virtual Reality Training System for Surgical Anatomy

(*continued*)

Table 4. (*continued*)

41	Eyestrain impacts on learning job interview with a serious game in virtual reality
42	Using Immersive Visualization Environments to Engage Students in Hands-On Learning
43	VR Application for Technology Education in a Public Library
44	Development of a model for explaining the learning outcomes when using3D virtual environments in informal learning settings
45	DAVEE : A Deaf Accessible Virtual Environment for Education
46	Keep the Ball Rolling : Designing Game-Based Tangible VR for Spatial Penetrative Thinking Ability
47	Virtual Reality Fire Disaster Training System for Improving Disaster Awareness
48	Usability of gamified knowledge learning in VR and desktop-3D
49	Supporting Anatomy Education with a 3D Puzzle in a VR Environment - Results from a Pilot Study
50	LIVE: the Human Role in Learning in Immersive Virtual Environments
51	Virtual Reality as a Means of Teaching Contemporary Chemistry
52	Do Low Cost Virtual Reality Devices Support Learning Acquisition?: A comparative study of two different VR devices
53	To Survive in a CBRN Hostile Environment: Application of CAVE AutomaticVirtual Environments in First Responder Training
54	Developing a virtual reality game for manufacturing education
55	Juggling in VR: Advantages of Immersive Virtual Reality in Juggling Learning
56	Virtual Reality Forge: Pattern-Oriented Authoring of Virtual Reality Nuggets
57	Investigating Representation of Text and Audio in Educational VR using Learning Outcomes and EEG
58	Evaluating Virtual Reality as a Medium for Vocational Skill Training
59	The Difficulties and Countermeasures of Applying Virtual Reality to Industrial Design Education
60	Improving Accessibility to Intangible Cultural Heritage Preservation Using Virtual Reality
61	Research on 3D Virtual Training Courseware Development System of Civil Aircraft Based on Virtual Reality Technology
62	Virtually the Same Experience?: Learning from User Experience Evaluation of In-vehicle Systems in VR and in the Field
63	Virtual Reality and Collaborative Interdisciplinary Work in the Development of Competences
64	The Effect of Presence and Appearance of Guides in Virtual Reality Exhibitions
65	Multi-user Virtual Training Assistant for Maintenance on Energized Medium-Voltage Lines
66	Embodying the homunculus in virtual reality: creating an interactive laboratory procedure for brain mapping and visualization
67	Design and Implementation of Virtual Simulation Experiment for Generation, Transmissionand Application of Electric Power
68	Visual Technologies for Urban Design Competences in Architecture Education
69	The Effects of VR Environments on the Acceptance, Experience, and Expectations of Cultural Heritage Learning
70	Towards Designing Agent Based Virtual Reality Applications for Cybersecurity Training
71	Blended Learning Supports in Chinese Architectures with 360 Degree VR Contents and 3D Modelling VR Contents
72	How Important is Immersion for Learning in Computer Science Replugged Games?
73	Design and Evaluation of a VR Training Simulation for Pump Maintenance
74	X-Road: Virtual Reality Glasses for Orientation and Mobility Training of People with Visual Impairments

(*continued*)

Table 4. (*continued*)

75	The Constructing and Application Case of Online Virtual Exhibits Arrangement System for Museum Learning
76	A Virtual Reality Training Tool to Improve Weight-Related Communication Across Healthcare Settings
77	Virtual Simulation of Palletizing Training for Industrial Robots
78	Multi-User System for Virtual Interaction of a Pasteurization Process
79	Virtual simulation of glue coating training for industrial robots
80	How to VizSki: Visualizing Captured Skier Motion in a VR Ski Training Simulator
81	Web-Portal-Based Repurposing of VR Scenarios for TEFL Applications

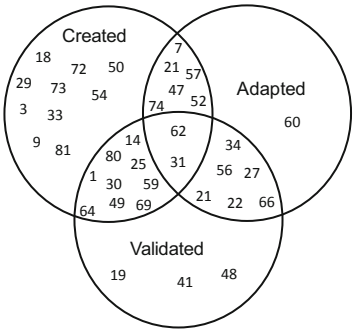

Fig. 3. Studies with complex evaluation grouped by classification regarding the kind of data collection instrument.

The created instruments usually were based on interviews, experiment observation, and use of experiment data, usability questionnaires, and learning outcomes surveys. Among the validated or adapted instrument, we can cite:

- Presence questionnaires: Witmer Presence [34], Igroup presence questionnaire[6], Multimodal Presence Scale [21] and ITC-Sense of presence inventory [19], a presence questionnaire adapted from [27] and translated in french;
- Usability and technology acceptance questionnaires: TAM [6], System Usability Scale [4], UTAUT [32], USE [20], AttrackDiff questionnaire[7], Depth of presence [30];
- Flow and discomfort questionnaires: Flow Short Scale [9], Visual Discomfort [35] and Simulator Sickness Questionnaire[14]; and
- Other tools: NASA-Task Load Index questionnaire [11], User Experience Questionnaire [18], a questionnaire extracted from [26], a questionnaire to measure the uncanny valley effect[13], an adaption from the questionnaire presented by [5], Think Aloud method [22] and Delphi Method to list the application requirements.

[6] http://www.igroup.org/index.php.
[7] http://attrakdiff.de/index-en.html.

Several studies used the following statistical tools to validate the adapted or created questionnaires: structural equation modeling, Cronbach Alpha test, double-sided T-test, and one study validated the data through scientific software[8]. It is also noted that some authors used other statistical methods, such as the Friedman test, independent sample T-test, and Wilcoxon signed-rank test, to discuss the obtained results.

We also collected the main statement presented as the conclusion of the study expressed regarding the experiment effects. First, it is important to point out that not every study is assertive about its outcome. However, most studies reports getting positive effects from the conducted evaluation: 4, 5, 12, 17, and 18. Study number 14 has also presented positive effects, yet the authors pointed out a difficulty in learning how to manipulate the new system interface.

5 Discussion

To better understand how educational software applications based on VR have been evaluated, we performed a systematic review in which 81 articles were analyzed. This study aimed to answer the five research questions (RQ), listed in Sect. 3, through the data collection step using a form. The questions are answered in the following subsections.

(RQ1 - RQ2) Based on the data presented in Table 2, the presence of studies employing a validated method is limited, whereas the typical approach is to adjust some method or create its own. In addition to the created methods, we can point to three approaches used in the analyzed studies:

1. adapted methods, that somehow follows the theory defined in HCI;
2. methods that use statistical analysis and hypothesis testing; and,
3. other approaches, such as design method and mixed-method.

A method must be based on a well-defined theory to provide reliability. More discretion is needed when using results from articles that created their evaluation without proper validation: studies without a well-defined method do not provide sufficient reliability to be generalized to other situations due to the lack of internal and external qualities [28].

The use of an adapted method, or at least structured in HCI and experimentation theories, can provide better results. Well-defined methods set up tasks, workgroups, questionnaire appliances, and data analysis, but only a few studies validate the collected data and point out outcome limitations. Thus, this also affects their reliability. There were only five articles that have validated their methods and tools by statistical analysis.

The lack of a well-defined experimentation structure can light problems regarding data validation. For instance, in studies 4 and 23, the users answered the same questionnaire before and after the intervention stage. In this case, the user can influence the results better or worse by remembering the previous

[8] https://www.ibm.com/br-pt/analytics/spss-statistics-software.

answer or question. This way, the results may reflect the person's memory, not the real intervention effects.

Despite the reasons and importance regarding the experimentation design, only one study has formally used them. The lack of an explicit definition could mean that the authors are unaware of it or do not take it for granted. This conclusion may also be extended to mixed methods since we noted studies that used more than one research approach and were not necessarily grounded upon a theoretical framework.

(RQ3) We can point out the lack of specific instrumentation to evaluate a VR application. The authors prefer to use usability, technology acceptance, and presence questionnaires to assess the new application. These are standard methods that can be used in general applications, not only VR-based.

This result is correlated to a previous systematic review, which had the objective of assessing how AR applications were evaluated with users [7]. Similar tools were found, such as usability assessments, interviews, and the measurement of tasks and data sensors. Even though the interactions with these domains—AR and VR—are different, they are analyzed nearly with the same methods listed in this article. Most of the methods listed by [7] were also not validated. Thus, they present an informal evaluation. Nevertheless, new research can change the scenario towards the tools' validation since its publication.

The lack of specific tools for performing VR and AR application evaluations may be due to the complexity of its interactions besides the required resources. Notwithstanding, 33 studies employed a more complex evaluation design composed of more than one data collection tool to perform such evaluation. Although it was not selected to further analysis due to the exclusion criteria, a study found during the search of this systematic review presents an evaluation measuring the application flow using the user's physiological data [3]. A similar approach was used by studies 22 and 30, which used EEG data to analyze some perspective of user experience. Thus, methods based on physiological sensors may be a possible solution to evaluate a VR application. However, it needs more resources to evaluate the system using this type of tool.

(RQ4) We classified the evaluation process of studies based on the number of data collection tools they had used: simple (i.e., used only one tool); and complex (i.e., used more than one instrument). The most common practice in studies belonging to the simple evaluation group is creating their instrument, which the only one that proceeded with a validation process of an adapted tool. As with all new methods, extra caution is required when analyzing and drawing conclusions since the experiment's internal quality is unsure.

The lack of support theories can not provide sufficient reliability; that is, there is no assurance when relating the variables and the interpretation. Thus, the analysis is based on the authors' not declared assumptions, hypothesis, and conceptions [28] and may not have sufficient formal pieces of evidence to support the analysis. In brief, the results will be limited and not accurately reflect the effects of the intervention.

It is also observed the choice for creating tools in complex evaluation studies. Figure 3 illustrates that most studies built their experiment using both created and validated tools. In contrast, few studies used an adapted tool composing the research - and only one adapted tool was validated. The use of non-validated tools among validated tools can provide better reliability of the interpretations and analysis - despite the discussion limitations due to lack of reliability. This article mentions that data supplied by sensors was also classified as a created tool, yet it may present accurate data. The qualitative tools were also classified as created tools in most cases. However, we can cite the Think Aloud method and Delphi method to develop the application requirements in two separate studies.

(RQ5) The predominant research classification signs the methodological principles considered by the studies, such as quantitative, qualitative, or even a combination of both approaches. An evaluation may use different tools (e.g., data collection instruments) accordingly to its paradigm, as observed in the results obtained through this study. Still, the tools are acknowledged through a predominant perspective or a specific method involving both aspects based on a mix-method theory. Regarding the research design, this study reveals the quantitative approach as favorite, with 38 articles, followed by some combined approach (13), and the least likely the qualitative approach (9).

6 Threats to Validity

The factors that can threaten this study's validity are presented following the concepts defined by [28]: construct, internal, and external validity. Construct validity involves the correct interpretation of concepts and if the results obtained are relevant or not. Despite evaluation concepts are well defined, there is a problem regarding the VR general concept: we eliminated studies that have not used a specific VR interface such as Oculus VR[9] and Google Cardboard[10]. This way, some relevant studies may be eliminated.

Internal validity implies that the relation between two desired variables, A and B, is consistent and not a measurement error. The limited amount of primary studies analyzed is a limitation of this study. We only used one source in our research, ACM Digital Library, and more sources would provide us a better broader understanding. Moreover, there was only one extractor, and another person did not validate the studies to find different assumptions and results.

External validity stands for the degree to which the results and interpretation can be generalized to other similar scenarios. The use of only one scientific repository is a concern about external validity. As noted before, more studies would be possible to find more accurate results. However, the results from this study provide a compilation of approaches and methods to evaluate a VR educational application. We also point out issues concerning the analyzed experiments, which serve as a guideline to help with setting evaluation in further research.

[9] https://www.oculus.com.
[10] https://arvr.google.com/cardboard/.

7 Conclusion

The obtained results show that most studies conduct evaluations based on their crafted procedures. That is, researchers created their methods or tools to investigate the effect of the intervention. Most of the analyzed studies have used adapted tools such as usability, user experience, flow questionnaires, task load questionnaire, the acceptance level of technology questionnaires, and other tools summarized. Moreover, many studies have used a method based on HCI theory related to experiment conduction. In contrast, only a few have handled more rigorous experimental designs, such as the Quasi-Experiment method and analysis through statistical hypothesis testing. Lastly, most studies employed quantitative approaches to synthesize and analyze the data to answer the research questions and hypotheses defined rather than qualitative or combined methods.

Although some studies reported a positive effect on the intervention of a VR educational application, more discretion is required to compare and generalize those results. The observed lack of support theories in many reviewed cases suggests high risks of evaluation flaws affecting data integrity, analysis, and, ultimately, the study's reliability.

Finally, we strengthen the relevance of using well-defined support theories in all stages of an evaluation. As future work, we propose expanding this research by exploring other fields such as VR in health and analyzing other scientific repositories.

Acknowledgments. This project has partially supported by Huawei do Brasil Telecomunicações Ltda (Fundunesp Process # 3123/2020), and CAPES.

References

1. American Educational Research Association, American Psychological Association, N.C.o.M.i.E.: Standards for Educational and Psychological Testing. Amer Educational Research Assn (2014)
2. Bach, C., Scapin, D.L.: Obstacles and perspectives for evaluating mixed reality systems usability. In: Acte du Workshop MIXER, IUI-CADUI, vol. 4. Citeseer (2004)
3. Bian, Y., Yang, C., Gao, F., Li, H., Zhou, S., Li, H., Sun, X., Meng, X.: A framework for physiological indicators of flow in vr games: construction and preliminary evaluation. Pers. Ubiq. Comput. **20**(5), 821–832 (2016)
4. Brooke, J., et al.: Sus-a quick and dirty usability scale. Usab. Eval. Ind. **189**(194), 4–7 (1996)
5. Csikszentmihalyi, M., Larson, R.: Validity and reliability of the experience-sampling method. In: Flow and the Foundations of Positive Psychology, pp. 35–54. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_3
6. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q., 319–340 (1989)
7. Dünser, A., Grasset, R., Billinghurst, M.: A survey of evaluation techniques used in augmented reality studies. Human Interface Technology Laboratory New Zealand (2008)

8. Elmore, P.B.: Reporting standards for research publications. Couns. Outcome Res. Eval. **1**(2), 19–29 (2010)
9. Engeser, S., Rheinberg, F.: Flow, performance and moderators of challenge-skill balance. Motiv. Emot. **32**, 158–172 (2008). https://doi.org/10.1007/s11031-008-9102-4
10. Greene, J.C., Caracelli, V.J., Graham, W.F.: Toward a conceptual framework for mixed-method evaluation designs. Educ. Eval. Policy Anal. **11**(3), 255–274 (1989)
11. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 904–908. Sage Publications Sage CA, Los Angeles (2006)
12. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. Manag. Inf. Syst. Q. **28**, 75–105 (2004)
13. Ho, C.C., MacDorman, K.F.: Measuring the uncanny valley effect. Int. J. Soc. Rob. **9**(1), 129–139 (2017)
14. Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G.: Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int. J. Aviat. Psychol. **3**(3), 203–220 (1993)
15. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele Univ. **33**(2004), 1–26 (2004)
16. Kitcher, P.: The Advancement of Science: Science without Legend, Objectivity without Illusions. Oxford University Press, Oxford (1995)
17. Kuhn, D., Pearsall, S.: Developmental origins of scientific thinking. J. Cogn. Dev. **1**(1), 113–129 (2000)
18. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89350-9_6
19. Lessiter, J., Freeman, J., Keogh, E., Davidoff, J.: A cross-media presence questionnaire: the itc-sense of presence inventory. Pres. Teleoper. Virtual Environ. **10**(3), 282–297 (2001)
20. Lund, A.M.: Measuring usability with the use questionnaire12. Usab. Interface **8**(2), 3–6 (2001)
21. Makransky, G., Lilleholt, L., Aaby, A.: Development and validation of the multimodal presence scale for virtual reality environments: a confirmatory factor analysis and item response theory approach. Comput. Human Behav. **72** (2017). https://doi.org/10.1016/j.chb.2017.02.066
22. Nielsen, J.: Thinking aloud: The# 1 usability tool. Nielsen Norman Group 16 (2012)
23. Rauschenberger, M., Schrepp, M., Pérez Cota, M., Olschner, S., Thomaschewski, J.: Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (ueq). example: Spanish language version. Int. J. Interact. Multimedia Artif. Intell. (2013)
24. Roberts, P.G., Guyver, P., Baldwin, M., Akhtar, K., Alvand, A., Price, A.J., Rees, J.L.: Validation of the updated arthros simulator: face and construct validity of a passive haptic virtual reality simulator with novel performance metrics. Knee Surg. Sports Traumatol. Arthrosc. **25**(2), 616–625 (2017)
25. Ruthenbeck, G.S., Reynolds, K.J.: Virtual reality for medical training: the state-of-the-art. J. Simul. **9**(1), 16–26 (2015)
26. Rutten, N., Van Joolingen, W.R., Van Der Veen, J.T.: The learning effects of computer simulations in science education. Comput. Educ. **58**(1), 136–153 (2012)
27. Schubert, T., Friedmann, F., Regenbrecht, H.: The experience of presence: factor analytic insights. Pres. Teleoper. Virt. Environ. **10**(3), 266–281 (2001)

28. Shadish, W.R., Cook, T.D., Campbell, D.T., et al.: Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell. Houghton Mifflin, Boston (2002)
29. Sharp, H., Rogers, Y.J.P.: Design de Interação: Além da interação humano computador. Bookman (2005)
30. Slater, M., Usoh, M., Steed, A.: Depth of presence in virtual environments. Pres. Teleoper. Virt. Environ. **3**(2), 130–144 (1994)
31. Sommerville, I.: Engenharia de software. PEARSON BRASIL (2011). https://books.google.com.br/books?id=H4u5ygAACAAJ
32. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. MIS Q., 425–478 (2003)
33. Villagrasa, S., Fonseca, D., Durán, J.: Teaching case: applying gamification techniques and virtual reality for learning building engineering 3d arts. In: Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 171–177 (2014)
34. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: a presence questionnaire. Presence **7**(3), 225–240 (1998)
35. Zeri, F., Livi, S.: Visual discomfort while watching stereoscopic three-dimensional movies at the cinema. Ophthal. Physiol. Opt. **35** (2015). https://doi.org/10.1111/opo.12194