UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ARTUR BECKER

The Landscape of XR Evaluation: Tertiary Review and Visualizations

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Profa. Dra. Carla M. Dal Sasso Freitas

Porto Alegre
March 2022

CIP — CATALOGING-IN-PUBLICATION

Becker, Artur

The Landscape of XR Evaluation: Tertiary Review and Visualizations / Artur Becker. – Porto Alegre: PPGC da UFRGS, 2022.

118 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022. Advisor: Carla M. Dal Sasso Freitas.

1. Evaluation. 2. Virtual reality. 3. Augmented reality. 4. Mixed reality. 5. Systematic review. 6. Tertiary review. 7. Information visualization. I. Freitas, Carla M. Dal Sasso. II. Título.

ABSTRACT

Extended reality (XR) applications - encompassing Virtual Reality, Augmented Reality, and Mixed Reality - are finding their way into multiple domains at an accelerating pace. Each area has different motivations for employing XR and various criteria for evaluating XR. Surveys from several fields described XR applications and how they are evaluated in that field. However, there is not always a clear definition of what XR is for each particular area, and when there is, it might differ substantially from the other areas. This lack of consensus on a definition makes it hard to compare XR research efforts across areas and learn from them. Through a tertiary systematic literature review, We surveyed 81 surveys published from several fields to form a comprehensive summary of the current state of XR research that deals with the evaluation of XR applications. Our survey is founded on understanding (i) how is XR defined? (ii) why is XR employed? (iii) how is XR evaluated? (iv) what are the main criticisms and future research paths outlined by the surveys? (v) how good are the surveys? We present our findings through visualizations of the "evaluation landscape" in XR and describe the state of XR research in each of the ten categories we cataloged. We identify gaps in the definition of XR, as well as limitations in the current effectiveness-based research. We propose that future research in XR should build upon a solid taxonomy to depart from effectiveness research and into efficiency research - to understand not only textitif but also *how* XR achieves the desired outcomes.

Keywords: Evaluation. virtual reality. augmented reality. mixed reality. systematic review. tertiary review. information visualization.

O Panorama de Avaliação em XR: Revisão Terciária e Visualizações

RESUMO

Aplicações de realidade estendida (XR) - abrangendo Realidade Virtual, Realidade Aumentada e Realidade Mista - estão encontrando seu caminho em vários domínios em um ritmo acelerado. Cada domínio de aplicação tem diferentes motivações para empregar XR e vários critérios para avaliar XR. Diversas *surveys* em diferentes áreas descrevem aplicações de XR e como elas são avaliadas. No entanto, nem sempre há uma definição clara do que é XR para cada área específica e, quando há, pode diferir substancialmente daquela usada em outras áreas. Essa falta de consenso sobre uma definição torna difícil comparar os esforços de pesquisa em XR entre as áreas e aprender com eles. Por meio de uma revisão sistemática terciária da literatura, analisamos 81 artigos ti[po *survey* publicados em vários domínios de aplicação para construir um resumo abrangente do estado atual da pesquisa em XR que envolve avaliação de aplicações de XR. Nossa pesquisa é baseada no entendimento de (i) como XR é definida? (ii) por que XR é empregada? (iii) como XR é avaliada? (iv) quais as principais críticas e caminhos de pesquisas futuras delineados pelos artigos estudados? (v) quão boas são as *surveys*? Apresentamos nossos resultados por meio de visualizações do "panorama de avaliação"em XR e descrevemos o estado da pesquisa de XR em cada uma das dez categorias que catalogamos. Identificamos lacunas na definição de XR, bem como limitações na pesquisa atual relacionada com a eficácia de XR. Propomos que a pesquisa futura sobre XR deve ser realizada sobre uma taxonomia sólida partindo da pesquisa sobre eficácia em direção à eficiência - para entender não só *se* mas também *como* XR atinge os resultados desejados.

Palavras-chave: avaliação, realidade virtual, realidade aumentada, realidade mista, revisão sistemática, revisão terciária, visualização de informações.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

XR	Extended Reality
AR	Augmented Reality
VR	Virtual Reality
SLR	Systematic Literature Review
MDS	Multidimensional scaling

CONTENTS

1 INTRODUCTION

Evaluation – the systematic acquisition and assessment of information to provide useful feedback about some object (TROCHIM; DONNELLY, 2008) – is a fundamental part of extended reality (XR - a term that encompasses virtual reality, augmented reality and mixed reality) research. Through evaluation, we can find if XR applications are safe, effective, and comfortable, whether they are more efficient than other applications that do not involve XR at all, and how different types of XR, employing different technologies, compare to one another.

Evaluation can take many forms, depending on its type (formative or summative) and the discipline carrying it out: formative evaluation aims to improve the object of study; summative evaluation assesses the effects of the object of study (TROCHIM; DONNELLY, 2008; LAVIOLA et al., 2017). What information is acquired to perform each evaluation, and the means of acquiring it also vary, both in terms of study designs (such as randomized controlled trials versus informal user evaluations) and data acquisition methods (ranging from interviews and observation to clinical scales and brainwave measurements). The field of human-computer interaction (HCI) might, for example, evaluate the ease-of-use of an artifact with the goal of improving it by measuring the time it takes for a user to complete a task with the artifact, and how many mistakes they make along the way (LAVIOLA et al., 2017), i.e., formative evaluation of efficiency and task success. On the other hand, Psychology and Medicine research might evaluate XR-based interventions with the goal of assessing its effect on patient outcomes (CORBETTA; IMERI; GATTI, 2015; TURNER; CASEY, 2014), by measuring the symptoms of the patient before and after the intervention with an appropriate scale, i.e., summative evaluation of effectiveness.

In fact, evaluation of XR applications can be so all-encompassing as to consider characteristics of the system's hardware and software, how the system performs the task it is purported to accomplish (including its long-term effects on the people using it, such as in the psychological or medical treatment examples above), and how the users of the application perceive the system (LAVIOLA et al., 2017). Laviola et al. (LAVIOLA et al., 2017) equate the evaluation of XR systems to usability evaluation. Their broad definition of usability encompasses both the ease of use of the application, as well as its effectiveness in achieving its intended goals. Similarly, in general Interaction Design research, (SHARP, 2019) includes both effectiveness and utility within the scope of usability. In

contrast, the standardized definition of usability in software development proposed by the International Standards Organization (ISO, 2018) breaks it down into three components: effectiveness (in terms of objectives – as in tasks – achieved and number and magnitude of errors), efficiency and satisfaction. (EXPERIENCE, 2012), purposefully separates usability and utility, considering them two components of usefulness.

In order to understand the state of XR research, it is vital to understand how XR applications are evaluated. By doing so, we may reveal which aspect of XR is most important for each domain area and potentially identify gaps in the evaluation of certain aspects of XR in certain fields. One obstacle to study evaluation in XR research is the sheer amount of research as more diverse areas start to employ and evaluate XR applications. Bibliometric analyses in the fields of Medicine (HAN et al., 2020), Education (KARAKUS; ERSOZLU; CLARK, 2019), and Rehabilitation (HUANG et al., 2016), show a steady growth in the number of publications using VR, AR and MR in those domains. This makes it difficult to perform a cross-area systematic review of the studies directly. A way around this is to perform a tertiary review of the systematic reviews already published in the different areas - an approach used in the fields of code smells and refactoring and technical debt (LACERDA et al., 2020; RIOS; NETO; SPÍNOLA, 2018). While this strategy does not afford the same granularity as looking at the primary studies directly, we deem it appropriate for providing an overview of a multidisciplinary topic, as explained in the next section.

1.1 Why perform tertiary reviews

Systematic literature reviews (SLRs) are a means to obtain unbiased aggregates of evidence on a certain topic by locating all studies that are relevant to it. They are widely used in evidence-based medicine, and Kitchenham and Charters (2007) proposed an adaptation of the method to support evidence-based software engineering. The main difference of SLRs compared to ad-hoc surveys is that the former are thorough and fair - the search strategy must be auditable and repeatable. The core principles of the method are: (i) a defined review protocol stating the research question and methods; (ii) a defined, reported search strategy to allow for replication and assessment of thoroughness; (iii) explicit inclusion and exclusion criteria to decide whether to include a study in the review; (iv) a specified set of information to retrieve from each study, including quality criteria. The studies that are reviewed in an SLR are called “primary studies”. SLRs themselves

can be called secondary studies. Besides the summarization of evidence, other reasons to perform SLRs are to find potential gaps in research on a certain field, and to serve as a backdrop to introduce new research.

A special form of SLRs are Tertiary reviews, which can be employed when there are enough secondary studies on the topic of interest, and conducting an SLR would be too costly. The method to perform a tertiary review is the same as an SLR. Kitchenham et al. (2010) performed a tertiary review of software engineering SLRs, and found the number of such publications to be increasing, while also improving in quality. Other recent tertiary reviews addressed: code smells and refactoring (LACERDA et al., 2020), drawing from the large number of secondary studies on both areas and studying the relationship between them; and the field of technical debt (RIOS; NETO; SPÍNOLA, 2018), using the information gathered to evolve the conceptual model for technical debt. Most of these reviews use *some type* of visualization to present part of the results: ranging from bar charts showing the number of included papers per year (LACERDA et al., 2020; RIOS; NETO; SPÍNOLA, 2018), radar plots for exploring the occurrence of topics on the primary studies inside the surveys (RIOS; NETO; SPÍNOLA, 2018), to network graphs showing the relationship between concepts (LACERDA et al., 2020; RIOS; NETO; SPÍNOLA, 2018). However, these visualizations are limited in scope and also by the non-interactive nature of the print medium. Some reviews go one step further and provide interactive visualizations that allow for the exploration of the survey results across several dimensions as introduced in the next section.

1.2 Why visualize review results

Some reviews use similar visualizations to allow the reader to interactively explore the included studies. Chatzimparmpas et al. (2020a) studies the topic of enhancing trust in Machine Learning models. They developed a web application, the TrustMLVis Browser, for interactively visualizing the survey results as a grid of thumbnails – each representing one visualization technique – and filter controls. Details of each technique are available on demand by clicking on each thumbnail. For the filters, the primary studies were coded into several dimensions (such as which domain they pertain to, what ML model is used, what type of visual representation is employed). Similarly, Kerren et al. (2017) reviews data visualization techniques in Biology, supported by a web application named BioVis to visualize the included studies. The main difference between BioVis and TrustMLVis

is that the former uses a spatial layout to arrange the items according to similarity. This spatial layout is obtained via dimensionality reduction applying Multidimensional scaling (MDS) on several factors. For some of these factors, the Jaccard index was used as a metric of dissimilarity. Each individual technique is represented by a thumbnail. Details of each technique are available by clicking on the thumbnail - including a sorted list of the most similar publications. The techniques are described in terms of a "broad taxonomy", comprising biological data types, biological data properties and visualization tasks. This taxonomy can be used to filter the dataset. In both surveys (CHATZIMPARMPAS et al., 2020a; KERREN et al., 2017), new entries can be added by the public, via a form on the websites. The addition is not automatic – the paper and coding has to be vetted by the authors.

Several systems were created with the sole purpose of providing such interactive visualizations in a modular manner. SurVis (BECK; KOCH; WEISKOPF, 2016) is a tool created for the analysis and dissemination of literature collections, which displays manually curated publications with very flexible filtering capabilities. StArt (FABBRI et al., 2012) aims to support the entire systematic review process, including visualizing the publications. VOSviewer (ECK; WALTMAN, 2010), which is available as an off-the-shelf computer program, allows for the creation and visualization of bibliometric maps, typically showcasing terms that are present in the literature collection. The terms found are spatially arranged according to their co-occurrence. Finally, Eitan et al. (2021) allows for the creation of a graph rooted on a single paper: the nodes of the graph are the most similar papers to it, pooled from a large database. The tool is available online for public use[1]. We explore these systems in more depth in Section 2.2.

1.3 Contributions

We opted to perform a tertiary review of evaluation in XR research, based on the available systematic reviews. To the best of our knowledge, there is no secondary or tertiary review about evaluation in XR (VR, AR and MR) across all topics where XR research is currently conducted or employed. The closest work we could find was (SUH; PROPHET, 2018) which did analyze VR, AR and MR applications and their evaluation, but was restricted to the social sciences[2]. Our goal is to provide a current overview so

[1] *connectedpapers.com*

[2] This paper was included in our tertiary review

that the similarities and differences between the several areas making use of XR are made explicit, as well as the patterns, focuses and gaps in how evaluation is carried out in each of them. Moreover, such a tertiary study allows us to observe the quality of the systematic reviews. In software engineering, for example, it has been observed that "secondary studies are often incomplete, and not always well organised" (BUDGEN et al., 2018). We also contribute with visualizations of this landscape of evaluation in XR, and in doing so we explore the limitations of current systems for visualizing survey results.

The contributions of our work are:

- A review of how XR is evaluated in different application domains, including the outcomes and measures employed.
- An analysis of the consistency of virtual, augmented and mixed reality definitions used throughout the reviews.
- An overview of why and how XR is used.
- What are main future research paths and criticisms for each domain according to the surveys.
- A quality assessment of the reviews.
- Visualizations of the landscape of evaluation in XR, which go beyond showing papers by representing outcomes and measures reported in the studied surveys.

1.4 Organization of the Text

This dissertation is structured as follows. Next chapter (Ch. 2) explores existing tertiary reviews and ways to visualize survey results in more depth. Then, Chapter 3 describes the survey method we adopted for performing the tertiary study. Chapter 4 explains the design and implementation process of our customized interactive visualizations built for analyzing the evaluation landscape. Chapter 5 presents the results from the 81 papers we analyzed and classified into 10 categories. In Chapter 6, we analyze and discuss the findings in each category. Finally, in Chapter 7, we provide conclusions and draw comments on future work.

2 RELATED WORKS

In this section, we go through some works related to (i) tertiary literature reviews and (ii) visualization of surveys and literature collections.

2.1 Tertiary reviews

The practice of systematic literature reviews in software engineering and related areas is relatively new, with the first specific guidelines being published in 2004 and revised in 2007 (KITCHENHAM, 2004; KITCHENHAM; CHARTERS, 2007). Since then, their usage as a means to thoroughly aggregate and synthesize evidence has spread in the area: Budgen et al. (2018) has found 178 SLRs published between 2010 and 2015 in five major software engineering journals. In some cases, there are already a large number of SLRs in a given topic of software engineering to warrant a tertiary review. Tertiary systematic reviews follow the same process of SLRs, but focus on gathering pre-existing systematic reviews, surveys and mapping studies on the field. We look at some of these tertiary reviews to understand how they are performed – what types of surveys are included, what method is used, and what are their conclusions.

Kitchenham et al. (2010), after providing the first guidelines on SLRs for software engineering, performed a tertiary review of software engineering SLRs, from the date of publication of their guidelines, 2004, to 2008. This review extended the collection of SLRs from a previous paper, which came from a set of thirteen journals and conferences and spanned 2004 – 2007. In this tertiary mapping study, their intent was to create a catalog of published SLRs and mapping studies that could be easily navigated by undergraduate students and practitioners. The search was expanded to 4 digital libraries and two indexing systems. They ultimately included 53 reviews, and noted that most were published in the more recent years of the period searched. From these, 12 were deemed of interest to practitioners – most focusing on industrial case studies and surveys. However, the authors note that all 12 lacked quantitative results to substantiate possible recommendations for practitioners. Regarding the possible use of SLR by undergraduate students, analyzing the topics of the reviews, they noted that at the time they were nowhere near spanning the entire SE curriculum. Regarding the quality of the reviews the authors found it to be increasing on the more recent studies. Quality was assessed using the Database of Abstracts of Reviews of Effects (DARE) criteria – a set of criteria

to evaluate the completeness, and validity of systematic reviews[1].

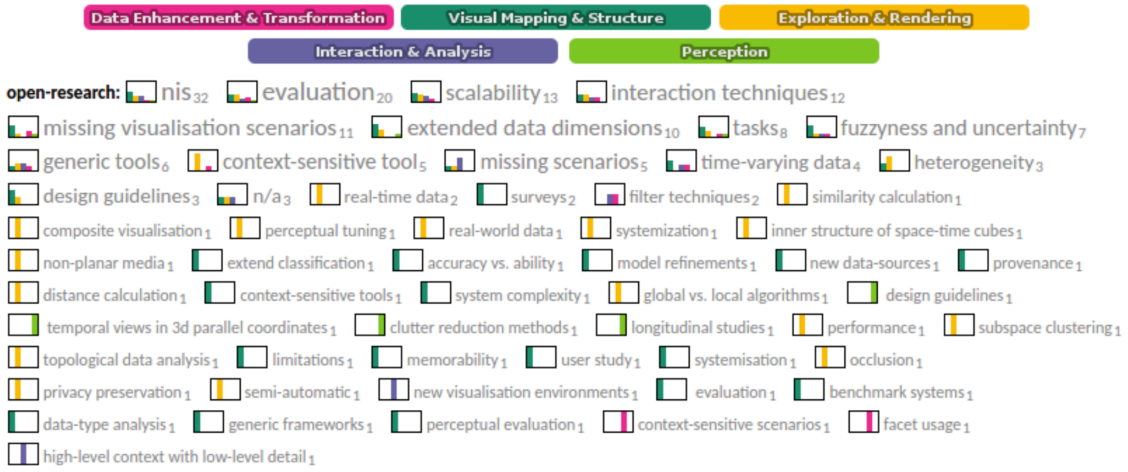
Budgen et al. (2018) performed a tertiary study to investigate how software engineering reviews are reported – and whether such reporting is conducive to the use of the SLR findings in software engineering teaching and practice. In the paper, they ultimately include 37 SLRs published between 2010 and 2015, gathered from five major software engineering journals. They consider that the two most important facets of reporting in reviews are: showing the provenance of conclusions and recommendations; and clearly portraying how the review was performed, in case others want to extend it. To investigate whether the included reviews provided adequate reporting, the authors assessed the reviews quality using the DARE criteria, and extracted lessons related to the scores, such as: (i) reviews should point out the inclusion and exclusion criteria on a dedicated table; and (ii) the process of applying these criteria should be illustrated as a diagram. The authors conclude that most reviews did perform thorough searches and were clear about their inclusion and exclusion criteria, but very few reviews actually reported material that might be useful for teaching or practitioners in the area (e.g., presenting a recommendation supported by the primary studies). Furthermore, they provide a checklist for software engineering SLRs to follow in order to improve their reporting.

Lacerda et al. (2020) reviewed SLRs, mapping studies and surveys on the topics of code smells and refactoring. They sought out to investigate what is known and what are the gaps in the current understanding of code smells and refactoring. The 40 reviews found by searching eight digital libraries spanned from 1992 to 2018. The reviews quality was assessed using the DARE criteria, with 80% of the reviews scoring between 3 and 4 (the highest score). The included surveys scored lower given that they usually employ a less formal protocol than SLRs and mapping studies. The authors outline the implications of the findings of their tertiary review for practitioners, instructors and researchers, as well as open challenges. The tertiary review made use of several visualizations – ranging from simple bar and pie charts to word clouds and Sankey diagrams analyzing the relationship between concepts found in the included studies.

Two other surveys that do not follow the systematic literature review approach but are of interest for us are by McNabb and Laramee (2017) and Chatzimparmpas et al. (2020b). Both are surveys of surveys (SoS) on the topic of Information Visualization and the usage of visualization to interpret ML models, respectively. McNabb and Laramee (2017) included 86 surveys, but since their SoS does not follow the SLR guidelines, the

[1] http://www.crd.york.ac.uk/CRDWeb/AboutPage.asp

Figure 2.1: Visualization of the open research topics in Visualization according to their category, using SurVis



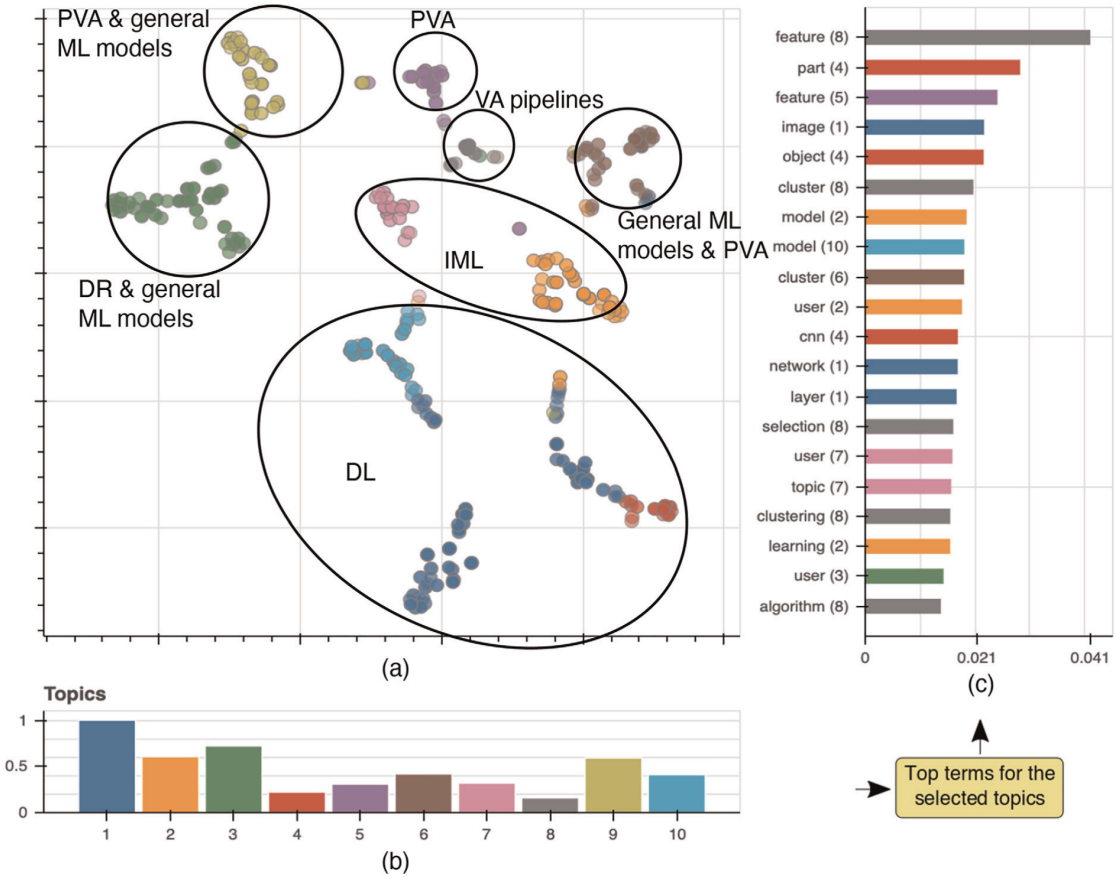
Source:(MCNABB; LARAMEE, 2017)

exact sources, queries and inclusion/exclusion criteria were not reported. The authors provide a classification of the surveys based on the information visualization pipeline (CARD; MACKINLAY; SHNEIDERMAN, 1999) as well as subjects discovered using SurVis. Also with the help of SurVis, the authors make available a list of topics still open to research according to their categorization (Fig. 2.1).

Chatzimparmpas et al. (2020b) is a SoS on the use of visualization for interpreting machine learning models. They searched for surveys using a broad array of keywords on 12 venues. The full queries, as well as inclusion and exclusion criteria were not reported. They ended up with 18 surveys from 2014 to 2018, and described (i) open challenges in the field, (ii) research subtopics, (iii) temporal and topical aspects of the primary studies included in the surveys. For the topic analysis, the authors processed each primary study and modeled the topics using latent Dirichlet allocation, which found 10 distinct topics. They plotted all primary studies in a two-dimensional chart obtained by using t-SNE for dimension reduction, and analyzed the topic clusters. They also visualized the primary studies and the surveys on a node-link diagram – showing what topics (each primary study was assigned a single topic) were explored by each survey (Fig. 2.2).

These two last examples illustrate the use of more advanced visualization techniques to help understanding literature collections. We will look at more methods for doing so – though not restricted to tertiary reviews – in the next section.

Figure 2.2: Visualization of papers on the field of ML model interpretation.



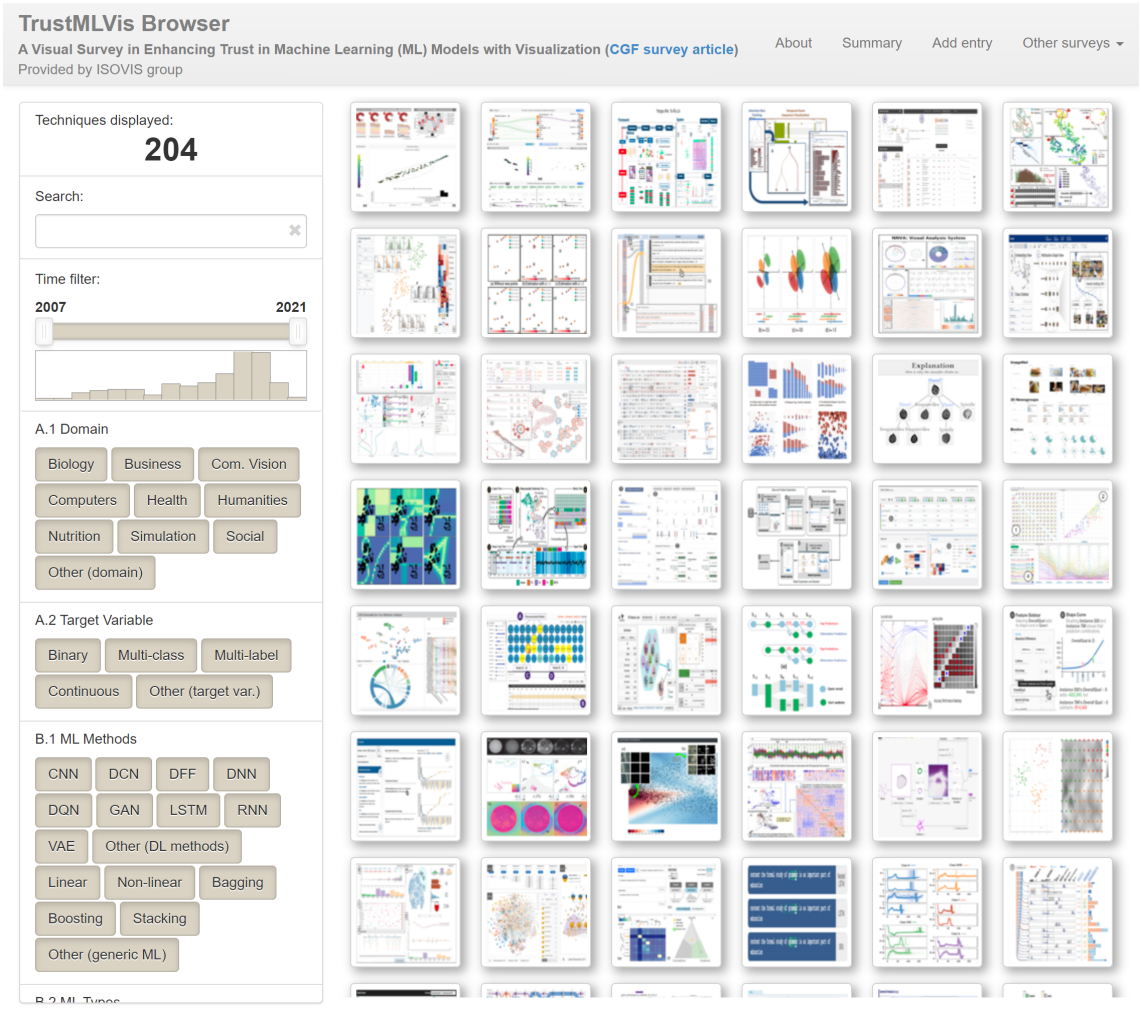
Source: (CHATZIMPARMPAS et al., 2020b)

2.2 Visualization of survey results

Although many review papers use some type of visualization to report their findings, most are limited to simple graphs, which the academic paper medium makes non-interactive by default. However, there are some reviews (and at least one SoS, as we described in the last section) that present the included papers as online-accessible interactive visualizations, normally supporting searching and filtering the collection of papers according to different properties of the papers. Schulz (2011) reports on Treevis.net, a "visual bibliography" for tree-visualization techniques, which today comprises more than 300 different entries. Each technique is presented as a small tile with an illustrative image, and details – including the link to the original publication – can be brought up on demand. Querying is supported on the full text of the publications, as well as filtering via three dimensions: dimensionality, representation, and alignment. Another survey (KEHRER; HAUSER, 2013) is based on the Treevis browser, but uses it to visualize techniques for multifaceted scientific data visualization – the main difference being the dimensions upon which the techniques can be filtered: data facet, technique, and main goal. Other visualizations follow a similar "browser-like" approach but include slightly more advanced filters, such as publication date, together with a histogram to show the frequency of each publication year (DUMAS; MCGUFFIN, 2014; LU et al., 2017), on the topics of finance visualization and predictive visual analytics, respectively. Several such visualizations come from the same research group, ISOVIS[2], from Linnaeus University in Sweden: text visualization (KUCHER; KERREN, 2015); sentiment visualization (KUCHER; PARADIS; KERREN, 2018); enhancing trust in ML through visualizations (CHATZIMPARMPAS et al., 2020a) (Fig. 2.3); and biological data visualization (KERREN et al., 2017). The latter is also unique among the browser-like visualizations because it uses the spatial arrangement of the technique thumbnails as a way to encode the similarity between techniques, which can also be calculated on-the-fly. Using MDS, the more similar techniques are displayed closer to each other. Finally, another slightly deviant example is by Schöttler et al. (2021). It displays the images of the techniques as well as basic data – title, tags and link – on tiles arranged in a mosaic-like disposition (thus evading the details-on-demand approach). The main similarity between all browser-like visualizations is that the theme of the surveys are visualization techniques, which lend themselves well for being represented as meaningful thumbnails.

[2] *https://cs.lnu.se/isovis/*

Figure 2.3: TrustMLVis Browser, an example of a browser-like visualization of survey results.



Source: *https://trustmlvis.lnu.se/*(CHATZIMPARMPAS et al., 2020a)

A second category of survey visualizations are the ones based on SurVis, (BECK; KOCH; WEISKOPF, 2016), a web-based visual analytics system that allows for the analysis of collections of publications. In the design process of the tool, the authors took into account two different roles: the curator of the collection and the reader, both of which have different goals and tasks. SurVis supports the reasoning process of both roles – the sensemaking loop of the curator and the foraging loop of the reader (PIROLLI; CARD, 2005). In the application, users can browse a list of publications (which pertain to the collection being analyzed), which can be filtered in several ways: through a timeline – which also provides an overview of the collection, through selecting words from a word cloud of keywords, from clusters generated for the collection, among others. In terms of visualization, bar charts are used to display the number of publications each year on a timeline, and this view is augmented with the number of citations for each publication (if any) as small tiles below the zero-line, with luminance encoding the citation number. Small sparkline charts are used alongside the text to indicate agreement between the filters applied. The system was evaluated by visualization experts, who were asked to use SurVis and provide feedback through a questionnaire, which had questions related to the requirements set for the system, but no explicit tasks to perform.

Some surveys that are available on SurVis for the interactive visualization of their bibliography are on: visualization of dynamic graphs (BECK et al., 2014); cartograms (NUSRAT; KOBOUROV, 2016); high-dimensional data (LIU et al., 2017); sparkline visualizations (BECK; WEISKOPF, 2017) ; analyzing scientific literature and patents (FEDERICO et al., 2017); and a survey of surveys in visualization (MCNABB; LARAMEE, 2017).

SurVis is fairly less reliant on the ability to meaningfully synthesize the included papers as images, since all papers are displayed in list form with identifiable information and an excerpt of the abstract – and thumbnails are optional. Survis also exposes and processes more information of the included papers – for example, by . creating clusters –, which can aid in the synthesis process of the review itself, as reported by McNabb and Laramee (2017).

As for off-the-shelf options, three solutions for visualizing literature collections are Connected Papers (EITAN et al., 2021), which allows for the creation of a graph from a single paper – and displays it as a node-link diagram: the nodes of the graph are the most similar papers to it, pooled from a large database. Similarity is computed as a function of the overlapping references and citations between each paper. The generated

network is represented as a force-directed graph, where each node represents a paper, the nodes opacity represents its publication date, and similar papers have increasingly thicker and darker links between them. Secondly, VOSviewer (ECK; WALTMAN, 2010), available as a computer program, allows for the creation and visualization of bibliometric maps, typically showcasing terms that are present in the literature collection, spatially arranged according to their co-occurrence. VOSviewer is unique compared to the other solutions presented in this section because it is the only one that focus on the *content* of the papers, rather than representing the papers themselves. Finally, StArt (FABBRI et al., 2012) is a tool to support the entire process of creating systematic literature reviews, from query creation to study inclusion/exclusion to synthesis. It allows for the creation of different visualizations based on the publications and their properties. Effectively, the system allows properties of the nodes (publications) to be represented as topology, a type of data operation according to a typology (NOBRE et al., 2019). Another feature of StArt is data mining: it automatically extracts the references of the papers, as well as calculates the similarity between included publications based on their abstracts.

2.3 Summary

In this chapter we briefly revise works on two topics closely related to our work: tertiary literature reviews and visualization of surveys and literature collections. We aimed at showing the importance of tertiary reviews to summarize the current knowledge about a topic and the methods they follow. We also use the results from this study to plan our tertiary review. Regarding visualization of surveys and literature collections, we provide an overview about what has been employed for communicating the results obtained from literature surveys. From this study, we came up with ideas for communicating our findings through different visualizations.

3 SURVEY METHOD

We performed a tertiary review of the literature to achieve a broad overview of XR research across multiple domain areas. We employed the methodology proposed by Kitchenham and Charters (2007), which has been widely adopted in software engineering (BUDGEN et al., 2018).

We seek to answer the following research questions:

- RQ1. How is XR evaluated?
  - Is evaluation formative or summative?
  - What outcomes, data collection methods and measures are used?
  - If evaluations are comparative, to what is XR compared?
  - Which designs are used in the evaluations?
- RQ2. How is XR defined?
  - What is the definition of Virtual, Augmented and Mixed reality in the study?
  - How are the XR applications described?
- RQ3. What is the motivation for using XR?
- RQ4. What are the main criticisms of primary studies and paths of future research for XR?
- RQ5. What is the quality of the current systematic reviews on XR?

3.1 Search process

Our search strategy included queries formed by the combination of three classes of terms, linked by AND operators: (i) XR terms, (ii) evaluation terms, (iii) SLR terms. Specifically, the terms on each of the three classes were: Virtual reality, Mixed reality, Augmented reality, Gesture; Evaluation, User study, Usability, User experience, Assessment; Systematic literature, Systematic review, Systematic mapping, Mapping study, Survey, Meta-analysis. We applied this search strategy in three digital databases, IEEE Xplore, ACM Digital Library and Science Direct. These databases were chosen for being large and relevant to the XR topic area. Titles, abstracts and keywords were searched. The full queries for each database can be found in the Supplementary material.

3.2 Inclusion and exclusion criteria

We included systematic reviews and meta-analyses that explicitly stated the search strategy employed (i.e., search strings and databases searched). Moreover, we only included papers published in the last decade: between 2010 and 2019.

Papers were excluded if they were:

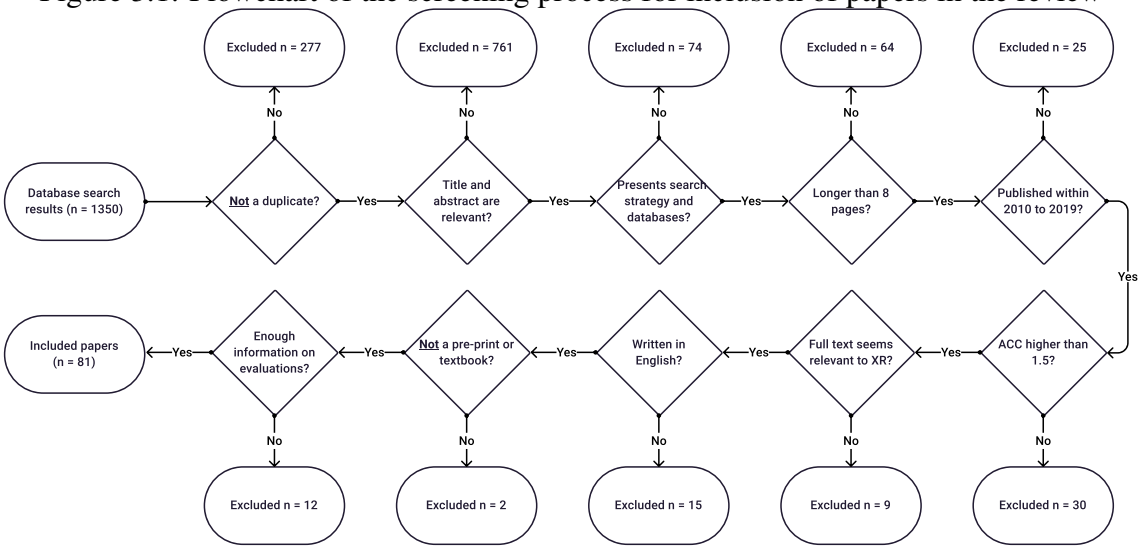
- Not about XR (i.e., some XR term is cited but no XR primary study is actually included)
- Not reporting information on the evaluations of the primary studies
- Not full-papers (i.e., shorter than 8 pages)
- Not written in English

Additionally, to focus on the most impactful papers, we excluded papers with an average yearly citation count (ACC) lower than 1.5 (citation counts were checked in Scopus or, if unavailable therein, Google Scholar, in July 2020).

3.3 Screening process

The queries returned 1,350 total results (268 from ACM Digital, 315 from IEEE Xplore, 767 from ScienceDirect). After the removal of 277 duplicates, all papers were screened for relevance to the topic based on their title and abstract, resulting in the exclusion of 761 papers. Because we were looking for systematic reviews, the methods sections of the remaining papers were examined to check if the search strategies and databases were explicitly stated. This excluded 74 papers. In the following step, we excluded papers that were shorter than 8 pages (n = 64), were published before January 1st, 2010 or after December 31, 2019 (n = 25), had ACC's lower than 1.5 (n = 30), or were not in English (n = 15). Additionally, 9 papers previously considered of ambiguous relevance to the topic in the first screening phase were excluded after a full-text skim. One study that was a journal preprint, and one retrieved result that was a textbook were also excluded. The remaining 93 studies were read in their entirety. After the full-text read, 12 studies were excluded for not providing enough information on the evaluations of the primary studies. Thus, 81 papers were ultimately included in this review.

Figure 3.1: Flowchart of the screening process for inclusion of papers in the review



3.4 Data extraction

We extracted the following information from the included papers:

- Metadata (Title, author, publication, year)
- Citation count (from Scopus or Scholar)
- Research questions or aim
- Type of review (SLR, Mapping study, Meta-analysis)
- Databases searched
- Years searched
- Number of included primary studies
- Number of XR primary studies (for broader SLR's that are not XR-exclusive)
- Quality score according to the DARE criteria [1]. Each of the following aspects is scored as 0, 0.5 or 1 [2], according to how well the criterion is met, and the final score is the sum of all four:
  - Are the review's inclusion and exclusion criteria described and appropriate?
  - Is the literature search likely to have covered all relevant studies?
  - Did the reviewers assess the quality/validity of the included studies?
  - Were the basic data/studies adequately described?

[1] http://www.crd.york.ac.uk/CRDWeb/AboutPage.asp
[2] See (KITCHENHAM; CHARTERS, 2007) for the rubric of each criterion

Additionally, to address our research questions, we extracted and coded the following points:

- What type of XR the review concerns (VR, AR, MR): this was coded from how the review self-identified (generally in its title, abstract or inclusion criteria). In the process, we extracted the review's inclusion criteria *ipsis litteris*.
- What is the definition of XR: if the review explicitly provided a definition of XR (generally in its introduction), we collected that definition *ipsis litteris*. Additionally, if references were provided for that definition, we collected them as well.
- What were the motivations cited for the use of XR: this was generally a dedicated segment in the review's introduction where the potential benefits of XR for the domain area are outlined, extracted *ipsis litteris*.
- How are the XR applications of the primary studies described: we extracted the passages of the review where the primary applications are described, if at all. For example, one review might state that among the primary studies' applications there were "immersive head-mounted displays and off-the-shelf video games". We concatenated all the passages where primary studies were described, *ipsis litteris*.
- What were the outcomes, methods and measures used in the evaluations of the primary studies. Below, we define what each of these terms means in our review.
- Study design: We extracted the different types of primary studies designs that appeared on each review. The types of designs are defined in the relevant section (see Sect. 5.3.3).
- Comparison: We extracted the different types of comparisons to XR made in the primary studies on each review, if any. The types of comparison are defined in the relevant section (see Sect. 5.3.4).
- Evaluation type: We extracted the types of evaluation in the primary studies that appeared on each review – formative and summative.

Regarding outcomes, methods and measures we clarify that **outcomes** are the main effects expected from the usage of XR. In some studies, these were literally called "outcomes"; in others, especially in narrative reviews, these were abstracted from the text. **Measures** are the specific means of gauging whether such outcomes are achieved and to what extent. Not all papers provided measures for the included outcomes. **Methods** are the nature of the measures used: for example, a measure can rely on self-report, such as an inventory of depression; or it can be based on observation, such as the time it took to

perform a task, as measured by an external observer.

Outcomes and measures were extracted *ipsis litteris* when possible, yielding a great variety of them. Some exceptions are studies where the outcome is implicit in the text, such as "Surgery training effectiveness" being the outcome of measures of surgery proficiency, "Learning effectiveness" being the outcome of tests of knowledge retention, or "System performance" being the outcome of measures of a system's tracking accuracy. After this first extraction, we reviewed all outcomes and measures and found some cases which were deemed appropriate – by both authors – for being merged together: examples are uniting the outcomes of "Upper extremity motor function" and "Upper limb motor function", or the outcomes of "Motion sickness", "Cybersickness" and "Simulator sickness".

The extraction of the different methods for each measure was not as open in that they stem from a predefined set of methods proposed by the ISO (ISO/IEC, 2016): User observation, Information from users and Inspection. We refer to these three categories as "Observation", "Self-report" and "Inspection". Additionally, we included a category of "Instrumented", for cases where the outcome is measured with the help of technological devices (such as heart rate sensors), and "System Log", when such measurements are provided by the XR system itself. During the coding process, we also found the need to include an "Expert review" category, for measures such as heuristic evaluation by usability experts.

During data extraction, each measure was assigned a method. When the method of a measure was unclear, a web search on the measure was conducted. This generally yielded a definition of the measure that allowed us to categorise it.

Finally, after completing the full-text read of all papers, they were categorized bottom-up using card sorting. Then, labels for each of the clusters that emerged were defined by both authors and recorded for each paper.

3.5 Data analysis

The extracted data was used to answer the research questions through narrative synthesis, tables and diagrams. RQ1 is addressed with a summary table of all studies in section 5. We categorized the papers according to their main topics and split the narrative synthesis (Section 6) into sections, one for each category. Each section presents the definition of XR and the description of the applications (RQ2), the main motivations for the

use of XR (RQ3), what the evaluations focused on (RQ1), and what are the criticisms and paths for future research (RQ4) for a category of studies.

4 VISUALIZATION OF THE SURVEY RESULTS

In this chapter we present the motivations for creating our own visualizations to support our analysis, based on the data we collected from the surveys – namely the network of outcomes and measures extracted from each paper.

4.1 Limitations of current approaches for visualizing our literature collection

As part of the visualization design process, we explored possible solutions from the existing literature, as well as off-the-shelf (See Sec. 2.2). Here we discuss their appropriateness and limitations for dealing with the network data we collected in our review.

SurVis (BECK; KOCH; WEISKOPF, 2016) allows curators to not only create the collection of publications records, but also augment it with data, similarly to coding the papers. The two main types of data that curators can add to the publications are keywords and citation data. Keywords are a way to provide structure to the publications – and being chosen and assigned by the curators following their own tagging taxonomy, they can be more consistent across the collection than the own publication author's keywords. Since the publications can have many keywords related to completely different subjects, such as methodology or paper type, keywords can be further divided into categories. Citations can be added as additional data, and the links are restricted between papers in the collection.

This curated data, together with SurVis interface design principle of making "everything selectable", affords a very flexible way to interact with and explore the publication collection: each selector (keyword, publication venue, author, etc.) is assigned a color, and all other entities on the UI get a sparkline barchart showing its agreement to that selector or group of selectors. This allows exploring the correlation between any dimensions of the publications: selecting an author would reflect in all the keywords sparklines if that keyword was employed by that author, and how many of the keywords appearances can be credited to that author (if the bar is full). Then, selecting a keyword would add a second bar to all sparklines, showing how often that keyword appears together with those entities.

BioVis (KERREN et al., 2017) also relies on manually curated data from the papers, to allow for the filtering and similarity calculations between publications. While, like most browser-like visualization systems, its UI is much more limited than SurVis, it stands out because of the use of 2D location to represent the similarity between all publi-

Table 4.1: Comparison of several characteristics of SurVis and BioVis, two systems used to visualize survey results

System	Items at once	Level of detail	Details on demand	Query by attribute	Query by item	Attributes visualized	Channels used	Attribute to attribute visualization	Data transformations
SurVis	∼5	High	Yes	Yes	Yes*	Up to 6	Color hue, Length	Co-occurence (many to 6)	Clustering
BioVis	∼100	Low	Yes (2 levels)	Yes	Yes*	5, dimensionally reduced to 2	2D position, additional link marks	Co-occurence (each to time) and to each other (indirect, via subsets)	Re-calculation of spatial layout, Similarity cut-off

* Attributes are displayed as part of the item “report card”

cations in the dataset, effectively creating a "visual landscape" of works of a certain field. We compare both SurVis and BioVis on Table 4.1.

While both solutions support the task of going from attributes of interest to publications, SurVis also supports exploring attributes from attributes, via the sparkline visualizations that appears on the side of every entity in the system. It is only possible to do that implicitly in BioVis, by selecting two attributes from the filter and looking at the effect this has on the set of publications shown: if the set did not change, the attributes are perfectly correlated; if the set shrunk, some papers that have the first attribute do not have the second one, but the ratio has to be imprecisely derived from the amount of publications shown. In BioVis, the only attribute that has a dedicated visualization of co-occurrence is publication date: the histogram of publication dates uses a stacked bar to show the number of publications on a given year from the full set versus the number of publications that agree with the selected filters – this way, all attributes and combinations of attributes can have its correlation with publication date visualized. In contrast, SurVis allows this to be done with up to six attributes to all other attributes (and publications). Figure 4.1 illustrates this difference on the interface of both systems.

Thus, discovering the relationship between attributes is possible on SurVis (though “hidden” behind one step of interaction: clicking on one attribute from the list), but is poorly supported in BioVis. Neither system is ideal to represent network data from the publications, if making the topology of the network explicit is important for the task at hand, as exemplified on 4.2. The main limitation of SurVis to visualize network data is that only approach (c) on Fig. 4.2 (a flat list of attributes, containing both attributes and links between attributes) would faithfully represent the network topology. If only raw attributes are used, their co-occurrence might not necessarily imply a link, it just means that both attributes appear on the same publication, as shown in Fig. 4.2 (a). However, a flat list of links is not as comprehensible as a dedicated visualization (e.g., a node-link

Figure 4.1: Comparison of (left) selecting an attribute on SurVis (LIU et al., 2017), which filters the set of publications, updates the histogram of publication dates and creates a sparkline bar chart showing the co-occurence of each attribute with the one selected, in this case, “interactive exploration”, and (right) the effects of selecting attributes on BioVis (KERREN et al., 2017), which filters the publications on the right and updates the histogram of publication dates to act similarly to a stacked bar chart.

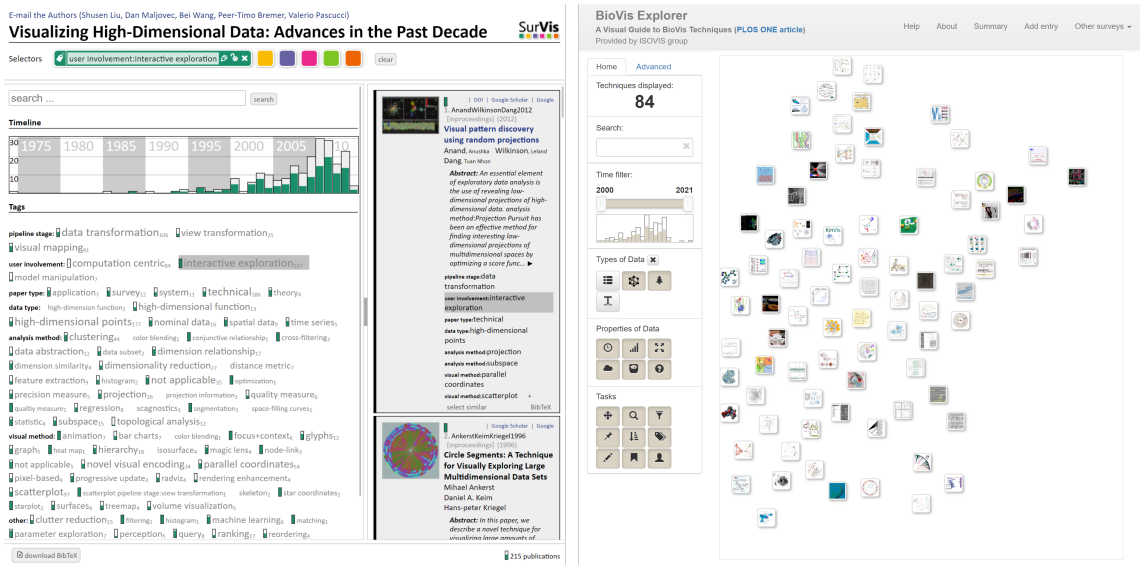
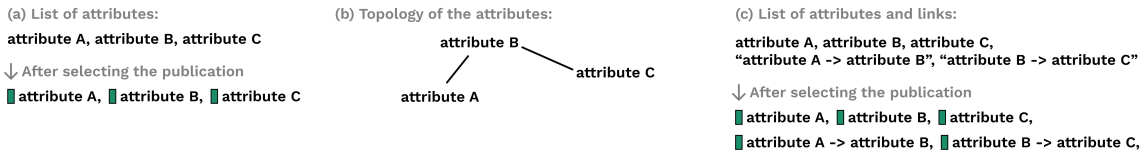


diagram or an adjacency matrix). Also, creating such pseudo-attributes that represent each link does not scale well for dense networks, where up to $n!$ additional attributes would need to be created.

Finally, VOSViewer (ECK; WALTMAN, 2010) does allow the visualization of terms co-occurrence among the papers. However, it suffers from the same limitations cited above for SurVis and BioVis: the co-occurrence is based on a paper unit of analysis, not a paper in a survey, like our manually extracted outcome and measure network. Also, when working with “term maps” on VOSViewer, the items on the visualization are the

Figure 4.2: Comparing different methods to visualize a network topology between attributes: (a) the attributes can be selected and links are implied by their co-occurrence - note that this doesn’t really represent the network topology, only co-occurrence on publications; (b) attributes and links between them represented as a node-link diagram; (c) attributes and links between attributes are “baked” in a single attribute list - this retains the network structure, but can result in $n + n!$ attributes on the list for fully connected graphs.



terms, while their co-occurrence is represented by the proximity of terms and also by links; gauging the overall co-occurrence of terms between two publications, however, is not supported. The main difference of VOSViewer is its focus on automatic data mining, as opposed to the intentionally “curated” nature of SurVis and BioVis. While the curated approach is much more time consuming, it can also produce more meaningful and reliable results when smaller datasets are being analyzed – “curation over automation” was one of the design objectives of SurVis.

4.2 Creating the visualizations of our survey results

In order to visualize the publications and the attributes we coded from each publication, we explored various types of visualizations, all implemented using the D3.js visualization library. We began by building three static visualizations:

- A node-link diagram of the network of outcomes and measures across all papers – effectively displaying a sum of all the individual graphs from each paper (Fig. 5.1).
- A scatter plot of the 81papers included in our SLR in "evaluation space" – taking all the attributes we coded from them to calculate their position, using MDS (Fig. 5.2, similar to the approach taken by Kerren et al. (2017)).
- A Sankey diagram of the "flow" of measures going through each category, outcome and nature (Fig. 5.3).

All these visualizations are presented in the next chapter to illustrate our survey results. While these static visualizations provide valuable insight into publications and the network of attributes (outcome and measures), they do not support the linked exploration of both publications and attributes at the same time. This is a vital task because it allows the viewer to understand both the overview of the research landscape as well as the specifics of each publication and category. Such visualization should support the following tasks:

- T1: Discover in which publications a certain outcome or measure appear.
- T2: Discern if the publication containing such attribute are from the same category or from several categories.
- T3: Discover what are the outcomes and measures present in a certain publication
- T4: Discover what are the outcomes and measures present in a certain category

- T5: Discover what measures are linked to a certain outcome
- T6: Discover whether this pairing of outcomes and measures vary from publication to publication or category to category
- T7: Discover how similar are the publications in terms of their outcomes and measures

There are several ways to support these tasks through a visualization. One straightforward way would be to treat our dataset as a layered network, with three levels: the top level are the secondary reviews, the mid-level are the outcomes, and the leaves are the measures, with no links within each set, only between sets. Then, a common visualization technique to use is the node-link diagram, like we already used for the outcomes and measures graph, but also including the publications. This would make tasks T1 and T3 simpler – requiring no explicit interaction – and simultaneous for all publications at once, at the cost of making the graph somewhat more cluttered. A drawback of this approach is that using simple link marks between outcomes and measures would result in information loss, since it would hide in which publication that measure is linked to that outcome – making it impossible to accomplish task T6.

Another approach is the adjacency matrix representation, but it is also challenged by our layered network structure – while it supports a large number of nodes without clutter, making out the *path* through more than a pair of nodes quite difficult (NOBRE et al., 2019). Also, due to the nature of the matrix and our data, there would be large empty sections, since no node from each level has links within its own level. A more appropriate tabular layout would be a quilt, which are specifically geared towards layered networks, preferably with no layer-skipping links, which is the case of our dataset. However, our network is not dense enough to take full advantage of tabular layouts in general, and quilts specifically also suffer from its very slanted degree distribution – meaning that, in some cases a single cell on the quilt would have to differentiate between 10+ origin publications.

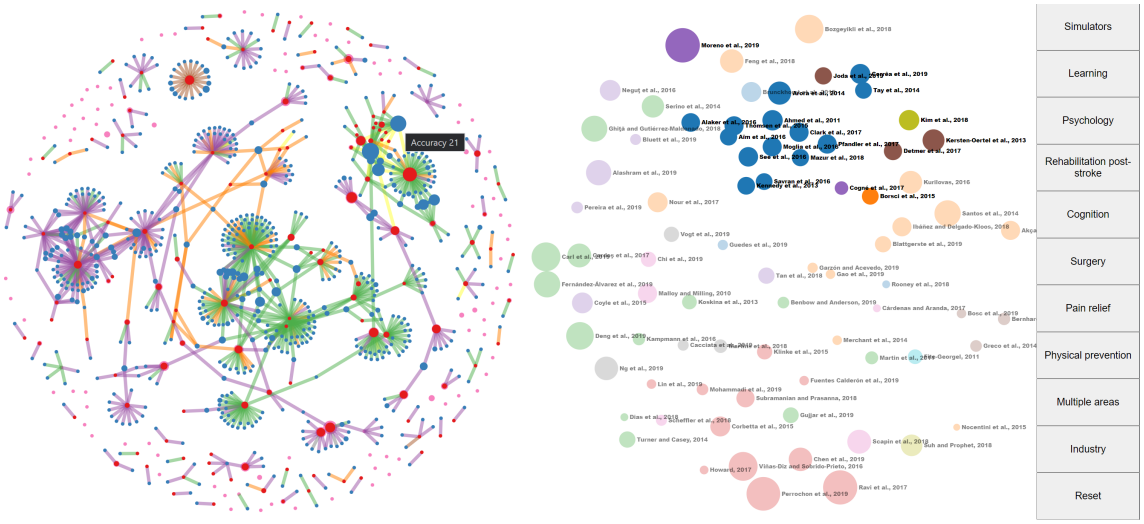
Finally, after analyzing the current solutions, reviewing possible layouts from the network visualization literature and making our own exploratory visualizations, we decided to build an interactive visualization that employs a juxtaposed view of the attribute network and the publications. The attribute network is displayed using a node-link diagram, and the publications on a scatter plot where one publication position is defined by their similarity to the others. In this way, all the tasks above can be accomplished – though with different interaction costs: tasks T5 and T7 can be accomplished by analyzing each view separately; the remaining tasks require some form of interaction. To support the

remaining tasks, we implemented the following interactions:

- T1 (Discover in which publications a certain outcome or measure appear): Hovering over an attribute highlights which publications have it.
- T2 (Discern if the publication containing such attribute are from the same category or from several categories): In this "attribute hover state", it is possible to distinguish which categories the publications that have the attribute pertain to.
- T3 (Discover what are the outcomes and measures present in a certain publication): Hovering a publication highlights its attributes, and shows a “report card” with more details about it. Additionally, clicking on a publication filters the network so that it can be examined in more detail.
- T4 (Discover what are the outcomes and measures present in a certain category): A category filter allows the filtering of the network to represent only the attributes from publications in that category.
- T6 (Discover whether this pairing of outcomes and measures vary from publication to publication or category to category): Clicking on an attribute filters its immediate neighborhood as well as the publications to show only the ones that have it. Then, hovering over the remaining publications highlights its attributes (e.g., the measures that appear on it).

The juxtaposed approach allows for the simultaneous visualization of 100 items and a network of 1,000 attributes. Through interaction, additional information about the publications (title, author, doi) can be accessed on demand. Several channels on both views are still used to represent other attributes, such as circle area for the amount of outcomes and measures on a certain publication, and node area for the amount of publications that have that attribute. Differently from our static MDS scatter plot, we used only the outcomes and measures as dimensions to calulate similarity, and thus positioning. Figure 4.3 shows a screen capture of the implemented visualization.

Figure 4.3: Interactive visualization showing both the 81 publications, arranged by similarity, and the network of outcomes and measures juxtaposed. The screen capture shows the effect of hovering over an attribute on the graph - what highlights the publications that include that attribute on the linked view. Category filters are available on the right-hand size to filter the graph, what is also possible by clicking on a publication



5 RESULTS

In this section, we present an overview of the 81 papers that complied with our inclusion criteria. Ten categories emerged during the analysis of the papers: Simulators, Learning, Psychology, Post-stroke Rehabilitation, Cognition, Surgery, Pain Relief, Physical Prevention, Multiple Areas, and Industry. Tables 5.1, 5.2 and 5.3 present the included studies, their categories, and the amount of primary studies and overlap between them.

5.1 Overview and Categories

In total, the reviews included 2673 primary studies[1], of which 2094 were unique. That means some of the primary studies appeared in more than one review. On average, our included papers contained 22.05%+-25.43% primary studies that were also included in at least one other review. When we consider only papers in the same category, the median overlap was 19.53%+-25.37%. Note that some of the reviews did not explicitly cite all primary studies, so it is possible this overlap is slightly larger. Two papers had all primary studies included in another paper: Martins et al. (2018) only reviewed one XR-related paper, also included in Ng et al. (2019)'s work; and Scheffler et al. (2018), which share 7 papers with Scapin et al. (2018). Davis, Nesbitt and Nalivaiko (2014) did not provide explicit citations to the primary studies, so we couldn't check its overlap.

Table 5.4 presents general information about the papers in each category. The median ACC of the reviews was 7.5 citations per year. The median publication year of the studies included is 2017, which hints at a growth in the number of XR papers published in recent years. The category with the lowest median publication year – not counting "Industry", which has only one paper – is "Simulators", at 2016, while the most recent is Physical prevention, at 2019. The median number of primary studies reviewed per paper is 28. The median span of the reviews is 11 years. The median DARE score was 3 out of 4. There was some convergence among the reviewed studies on what database was used to search for primary studies: PubMed was used in 41 reviews. On the other hand, publication venues were scattered: the top publication venue appeared only 5 times.

Most papers (69) included VR applications, while 21 included AR applications and only 5 included MR applications. VR is the most popular type of XR in every category except for Industry. VR is also the sole type of XR in three categories: Psychology,

[1] When a survey was not exclusively about XR, we only counted the primary studies that were XR related

Table 5.1: Included studies and the amount of overlapping primary studies – Simulators and Learning

Category	Citation	Author(s)	Primary studies	Overlapping studies	Percentage overlap	Percentage overlap within category
Simulators	(MAZUR et al., 2018)	Mazur et al., 2018	9	3	33.33%	33.33%
	(THOMSEN et al., 2015)	Thomsen et al., 2015	49	0	0.0%	0.0%
	(GUEDES et al., 2019)	Guedes et al., 2019	17	4	23.53%	23.53%
	(TAY; KHAJURIA; GUPTE, 2014)	Tay et al., 2014	14	3	21.43%	21.43%
	(CORRÊA et al., 2019)	Corrêa et al., 2019	145	5	3.45%	3.45%
	(ROONEY et al., 2018)	Rooney et al., 2018	7	1	14.29%	14.29%
	(AHMED et al., 2011)	Ahmed et al., 2011	21	0	0.0%	0.0%
	(KENNEDY; MALDONADO; COOK, 2013)	Kennedy et al., 2013	12	0	0.0%	0.0%
	(CLARK et al., 2017)	Clark et al., 2017	15	7	46.67%	46.67%
	(PFANDLER et al., 2017)	Pfandler et al., 2017	19	6	31.58%	31.58%
	(MOGLIA et al., 2016)	Moglia et al., 2016	36	1	2.78%	2.78%
	(ALAKER; WYNN; ARULAMPALAM, 2016)	Alaker et al., 2016	15	4	26.67%	26.67%
	(AÏM et al., 2016)	Aïm et al., 2016	9	3	33.33%	33.33%
	(BRUNCKHORST et al., 2015)	Brunckhorst et al., 2015	20	0	0.0%	0.0%
	(ARORA et al., 2014)	Arora et al., 2014	21	0	0.0%	0.0%
	(SEE et al., 2016)	See et al., 2016	31	1	3.23%	3.23%
	(SAVRAN et al., 2016)	Savran et al., 2016	26	0	0.0%	0.0%
Learning	(MERCHANT et al., 2014)	Merchant et al., 2014	65	0	0.0%	0.0%
	(SANTOS et al., 2014)	Santos et al., 2014	60	7	11.67%	10.0%
	(IBÁÑEZ; DELGADO-KLOOS, 2018)	Ibáñez and Delgado-Kloos, 2018	28	16	57.14%	50.0%
	(BLATTGERSTE; RENNER; PFEIFFER, 2019)	Blattgerste et al., 2019	52	2	3.85%	3.85%
	(BORSCI; LAWSON; BROOME, 2015)	Borsci et al., 2015	8	0	0.0%	0.0%
	(FENG et al., 2018)	Feng et al., 2018	15	0	0.0%	0.0%
	(NOUR et al., 2017)	Nour et al., 2017	3	0	0.0%	0.0%
	(GARZÓN; ACEVEDO, 2019)	Garzón and Acevedo, 2019	64	28	43.75%	42.19%
	(AKÇAYIR; AKÇAYIR, 2017)	Akçayır and Akçayır, 2017	43	19	44.19%	39.53%
	(KURILOVAS, 2016)	Kurilovas, 2016	33	6	18.18%	15.15%
	(NOCENTINI; ZAMBUTO; MENESINI, 2015)	Nocentini et al., 2015	2	0	0.0%	0.0%
	(BOZGEYIKLI et al., 2018)	Bozgeyikli et al., 2018	24	0	0.0%	0.0%
	(GAO; GONZALEZ; YIU, 2019)	Gao et al., 2019	15	1	6.67%	6.67%

Table 5.2: Included studies and the amount of overlapping primary studies – Psychology and Post-stroke rehabilitation

Category	Citation	Author(s)	Primary studies	Overlapping studies	Percentage overlap	Percentage overlap within category
Psychology	(BENBOW; ANDERSON, 2019)	Benbow and Anderson, 2019	46	22	47.83%	47.83%
	(CARL et al., 2019)	Carl et al., 2019	30	25	83.33%	83.33%
	(DIAS; BARBOSA; VIANNA, 2018)	Dias et al., 2018	1	0	0.0%	0.0%
	(GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018)	Ghiţă and Gutiérrez-Maldonado, 2018	13	0	0.0%	0.0%
	(FERNÁNDEZ-ÁLVAREZ et al., 2019)	Fernández-Álvarez et al., 2019	36	27	75.0%	75.0%
	(CARDOŞ; DAVID; DAVID, 2017)	Cardoş et al., 2017	11	10	90.91%	90.91%
	(KAMPMANN; EMMELKAMP; MORINA, 2016)	Kampmann et al., 2016	3	1	33.33%	33.33%
	(TURNER; CASEY, 2014)	Turner and Casey, 2014	30	21	70.0%	56.67%
	(GUJJAR et al., 2019)	Gujjar et al., 2019	1	0	0.0%	0.0%
	(KOSKINA; CAMPBELL; SCHMIDT, 2013)	Koskina et al., 2013	13	1	7.69%	7.69%
	(MARTIN et al., 2011)	Martin et al., 2011	1	0	0.0%	0.0%
	(DENG et al., 2019)	Deng et al., 2019	18	5	27.78%	27.78%
Post-stroke rehabilitation	(VIÑAS-DIZ; SOBRIDO-PRIETO, 2016)	Viñas-Diz and Sobrido-Prieto, 2016	25	12	48.0%	48.0%
	(KLINKE et al., 2015)	Klinke et al., 2015	2	0	0.0%	0.0%
	(CORBETTA; IMERI; GATTI, 2015)	Corbetta et al., 2015	15	11	73.33%	73.33%
	(CHEN et al., 2019)	Chen et al., 2019	12	1	8.33%	8.33%
	(CALDERÕN et al., 2019)	Fuentes Calderón et al., 2019	14	2	14.29%	14.29%
	(LIN et al., 2019)	Lin et al., 2019	4	1	25.0%	25.0%
	(PERROCHON et al., 2019)	Perrochon et al., 2019	11	1	9.09%	0.0%
	(MOHAMMADI et al., 2019)	Mohammadi et al., 2019	14	7	50.0%	50.0%
	(SUBRAMANIAN; PRASANNA, 2018)	Subramanian and Prasanna, 2018	5	2	40.0%	40.0%
	(RAVI; KUMAR; SINGHI, 2017)	Ravi et al., 2017	31	5	16.13%	12.9%
	(HOWARD, 2017)	Howard, 2017	113	16	14.16%	13.27%

Table 5.3: Included studies and the amount of overlapping primary studies – Cognition, Pain relief, Physical prevention, Multiple areas and Industry

Category	Citation	Author(s)	Primary studies	Overlapping studies	Percentage overlap	Percentage overlap within category
Cognition	(COGNÉ et al., 2017)	Cogné et al., 2017	60	1	1.67%	1.67%
	(ALASHRAM et al., 2019)	Alashram et al., 2019	9	1	11.11%	11.11%
	(NEGUŢ et al., 2016)	Neguţ et al., 2016	13	0	0.0%	0.0%
	(PEREIRA et al., 2019)	Pereira et al., 2019	3	1	33.33%	0.0%
	(COYLE; TRAYNOR; SOLOWIJ, 2015)	Coyle et al., 2015	3	1	33.33%	33.33%
	(BLUETT; BAYRAM; LITVAN, 2019)	Bluett et al., 2019	12	0	0.0%	0.0%
	(TAN; LEE; LEE, 2018)	Tan et al., 2018	2	1	50.0%	0.0%
	(MORENO et al., 2019)	Moreno et al., 2019	22	3	13.64%	4.55%
Pain relief	(SCHEFFLER et al., 2018)	Scheffler et al., 2018	7	7	100.0%	100.0%
	(CÁRDENAS; ARANDA, 2017)	Cárdenas and Aranda, 2017	2	0	0.0%	0.0%
	(CHI et al., 2019)	Chi et al., 2019	9	0	0.0%	0.0%
	(MALLOY; MILLING, 2010)	Malloy and Milling, 2010	11	1	9.09%	0.0%
	(SCAPIN et al., 2018)	Scapin et al., 2018	34	8	23.53%	20.59%
Physical prevention	(VOGT et al., 2019)	Vogt et al., 2019	16	2	12.5%	6.25%
	(NG et al., 2019)	Ng et al., 2019	22	7	31.82%	27.27%
	(CACCIATA et al., 2019)	Cacciata et al., 2019	9	0	0.0%	0.0%
	(MARTINS et al., 2018)	Martins et al., 2018	1	1	100.0%	100.0%
	(NEUMANN et al., 2018)	Neumann et al., 2018	20	5	25.0%	20.0%
Multiple areas	(KIM et al., 2018)	Kim et al., 2018	116	5	4.31%	0.0%
	(SUH; PROPHET, 2018)	Suh and Prophet, 2018	54	14	25.93%	0.0%
	(DAVIS; NESBITT; NALIVAIKO, 2014)	Davis et al., 2014	171	N/A	N/A	N/A
Industry	(FITE-GEORGEL, 2011)	Fite-Georgel, 2011	51	3	5.88%	0.0%

Post-stroke Rehabilitation and Cognition. In the next section we discuss the variations in this self-description and what are the theoretical underpinnings of XR in the included papers.

5.2 Definition of XR and the theoretical background of the surveys

Not all studies explicitly defined XR. This varied according to the paper category, with Simulators and Psychology having the lowest proportion of surveys providing a definition (around one quarter and half, respectively). Some possible explanations for this are: these two are the first and third largest categories in number of surveys, which might indicate that XR (in practice, VR) is a somewhat well-established concept in the field, and thus an explicit definition has become optional. However, this has the effect of allowing a reader to go through most surveys of the field not knowing what definition of Virtual Reality the authors subscribe to nor what the VR applications consisted of in terms of software and hardware. Among the papers that did define XR, not all provided references for it. We collected all the references used to define XR and present them in Table 5.4, split by category. In total, 66 different sources were cited to define XR. Simulators, Psychology and Pain Relief were the categories with the lowest ratio of XR-defining references, while Physical Prevention, Multiple Areas, Surgery and Learning had the highest ratios, around or above two thirds. The six papers most cited as definitions of XR were: Azuma (1997) (7 citations), Milgram and Kishino (1994) (6 citations), and tied with two citations each: Weiss et al. (2006), Cruz-Neira, Sandin and DeFanti (1993), Azuma et al. (2001), LaValle (2015).

Both most cited references are from the 90's. Azuma (1997) defines AR as any system that: (i) combines real and virtual; (ii) offers real-time interactivity, and (iii) the combination of real and virtual objects occurs in 3D. The authors use this definition to avoid tying AR to any specific technology (such as HMDs). They also cite Milgram and Kishino (1994) when comparing AR to VR, in which AR and VR are placed along the "virtuality continuum", as types of MR displays; another example of MR being Augmented Virtuality (AV). They define Mixed Reality systems according to three axes: Extent of World Knowledge, Reproduction Fidelity, and Extent of Presence Metaphor. Ultimately, if an MR application would be "all the way to the right" on the latter two, virtual and real objects would be indistinguishable, while the former adds the capability of actually interacting and enhancing the real world. When analyzing the definition of

XR technologies in our included papers, we notice a trend towards using the term VR to categorize fully synthetic, most often interactive worlds (although not necessarily fully immersive), and AR as the mixing of real and virtual objects. The MR label as well AR's sibling subclass, AV (MILGRAM; KISHINO, 1994), are seldom found in the analyzed surveys. Similarly, the axes of Extent of World Knowledge, Reproduction Fidelity, and Extent of Presence Metaphor are never used to define the XR applications, nor are they systematically employed as outcomes over which the applications are evaluated. A minority of papers do, however, evaluate presence, reproduction fidelity, and related outcomes separately, as shown in the next section.

5.3 Evaluation across XR

In this section, we present an overview of evaluation in XR. We try to answer the four parts of our first research question: "Is evaluation formative or summative? What outcomes, data collection methods and measures are used? If evaluations are comparative, to what is XR compared? Which designs are used in the evaluations?"

5.3.1 Outcomes, measures, and data collection methods

There was a great variety of outcomes and ways of measuring them in the included reviews. We catalogued 227 unique outcomes among the papers, and 649 unique measures. These are combined into 976 unique outcome-measure pairs. Additionally, we categorized each measure into one of six data collection methods (or "nature"):

- Observation: measures of user response recorded by an observer
- Self-report: measures that come from the user
- Instruments: measures recorded by an instrument, such as heart rate
- Inspection: measures that do not involve using the system directly, such as frame rate
- System log: measures recorded by the XR application itself, for example, in a simulator
- Expert Review: measures based on expert evaluation of the system

The top five outcomes, measures, and outcome-measure pairs are summarised in

	Count	ACC	Publication year	Studies reviewed	Start year	End year	Span	DARE score	Publication venue	Top database used	VR, AR, MR studies	Defines XR	Provides a reference for XR	References for XR (n)
All	81	7.5	2017	28	2004	2015	11	3	Computers in Human Behavior (5)	Pubmed (41)	69, 21, 5	49	34	
Simulators	17	6.57	2016	21	2002	2014	12	3.5	International Journal of Surgery (4)	Embase (13)	17, 2, 1	4	1	(PIMENTEL; TEIXEIRA, 1993) (1), (AZUMA, 1997) (1)
Learning	13	11.25	2017	33	2006	2016	9	3	Computers & Education	Scopus (7)	8, 7, 3	8	8	(AZUMA, 1997) (3), (LAVALLE, 2015) (2), (AZUMA et al., 2001) (2), (MILGRAM; KISHINO, 1994) (2), (GUO; YU; SKITMORE, 2017) (1), (CAUDELL; MIZELL, 1992) (1), (BOWMAN; GABBARD; HIX, 2002) (1), (AKÇAYIR; AKÇAYIR, 2017) (1), (CARMIGNIANI et al., 2011) (1), (HALE; STANNEY, 2015) (1), (HALLER; BILLINGHURST; THOMAS, 2007) (1), (SHARPLES et al., 2008) (1), (PERLMAN; SACKS; BARAK, 2014) (1), (FENG et al., 2018) (1), (CRUZ-NEIRA; SANDIN; DEFANTI, 1993) (1), (QUORA, 2018) (1), (WINN, 1993) (1)
Psychology	12	8.165	2018	29	1999.5	2015	12	3.5	Journal of Anxiety Disorders (4)	Psycinfo (8)	12, 0, 0	5	3	(FOX; ARENA; BAILENSON, 2009) (1), (GERARDI et al., 2008) (1), (GERARDI et al., 2010) (1), (ROTHBAUM et al., 2006) (1), (BOTELLA et al., 2007) (1), (RT ET AL., 2009) (1), (REPETTO; RIVA, 2011) (1), (PARSONS; RIZZO, 2008) (1), (BAÑOS et al., 2011) (1)
Post-stroke rehabilitation	11	4	2018	25	2009	2016	5	3.5	Tie	Pubmed (7)	11, 0, 0	8	6	(WEISS et al., 2006) (2), (MERIANS et al., 2002) (1), (BURDEA; COIFFET, 2003) (1), (LAVER et al., 2011) (1), (SCHULTHEIS; RIZZO, 2001) (1), (CRUZ-NEIRA; SANDIN; DEFANTI, 1993) (1), (WEISS; TIROSH; FEHLINGS, 2014) (1), (HINCKLEY, 2002) (1), (BOWMAN, 2005) (1)
Cognition	8	6.335	2018.5	19	2004	2015.5	11.5	3	Tie	Pubmed (6)	8, 0, 0	7	4	(PARSONS, 2012) (1), (RHEINGOLD, 1991) (1), (RIVA, 1997) (1), (GAMBERINI, 2000) (1), (SCHULTHEIS; HIMELSTEIN; RIZZO, 2002) (1), (KNIGHT; TITOV, 2009) (1), (MILLER; BUGNARIU, 2016) (1), (ROSE et al., 2001) (1), (RAND et al., 2005) (1), (LALONDE et al., 2013) (1), (KU et al., 2003) (1), (ELKIND et al., 2001) (1), (SLATER; WILBUR, 1997) (1)
Surgery	6	9.18	2017	65.5	1998	2016	14.5	2.75	Tie	Google Scholar (5)	2, 6, 0	5	4	(MILGRAM; KISHINO, 1994) (2), (MCCLOY; STONE, 2001) (1), (LOWOOD, 2015) (1), (FEINER, 2002) (1), (AZUMA, 1997) (1)
Pain relief	5	6	2018	21	2004	2011	10	3.5	Burns (2)	Psycinfo (4)	5, 1, 0	4	1	(HOFFMAN et al., 2001) (1)
Physical prevention	5	3	2019	16	2008	2016	8	3	Tie	Pubmed (4)	4, 2, 0	4	4	(SHERMAN; CRAIG, 2002) (1), (ANDERSON-HANLEY; ARCIERO; Snyder, 2011) (1), (BAÑOS et al., 2000) (1), (BOAS, 2012) (1), (MILGRAM; KISHINO, 1994) (1), (HOWARD, 2017) (1)
Multiple areas	3	13.5	2018	171	2009	2017	8	2	Tie	Scopus (2)	2, 2, 1	3	2	(KOLASINSKI, 1995) (1), (LEE; CHUNG; LEE, 2013) (1), (PRIBEANU; BALOG; IORDACHE, 2017) (1), (ROCHLEN; LEVINE; TAIT, 2017) (1), (WOJCIECHOWSKI; CELLARY, 2013) (1), (DUNLEAVY; DEDE; MITCHELL, 2009) (1), (ZENG; RICHARDSON, 2016) (1), (MILGRAM; KISHINO, 1994) (1), (AZUMA, 1997) (1), (LAVIOLA, 2000) (1)
Industry	1	19.44	2011	52	1998	2010	12	3.5	See (FITE-GEORGEL, 2011)	None	0, 1, 0	1	1	(AZUMA, 1997) (1)

[a] In case of surveys that are not exclusively abour XR, we only count the primary papers that were XR-related

Table 5.5. Each appearance of an outcome or measure on a survey paper contributes 1 towards its total, regardless of how many of the primary studies employed them.

Most of the outcomes (71%) appear in only one survey paper each. Most of the measures (82%) appear in only one survey paper each. Most of the outcome-measure pairs (91%) appear in only one survey paper each.

We attributed “Training outcomes” the 1st place and “Construct validity”, the 5th place based on their "popularity" in the “Simulators” category, which is the largest and one of the most homogeneous groups of survey papers. “Learning outcomes” are exclusive to the Learning category, and are an umbrella term used to refer to the way XR has helped improve learning. The other two, "Satisfaction" and "Usefulness" are characterized by being general enough so that they can span multiple categories.

The top two measures, "Time" and "Accuracy", are common objective measures that can be expected to span multiple categories. "Motion" is specific to the “Simulators” category (except for one appearance in “Surgery”), and is a measure of the quality of the users’ motion during the simulation, as recorded by the simulator. The 4th and 5th places belong to measures of time and accuracy in "real" activities completely detached from the system (i.e., how performance in a real surgery was affected by using the XR application on a separate occasion, for training), and thus they are differentiated from the simple “Time” and "Accuracy" measures.

The top outcome-measure pairs all consist of a combination of the top 5 outcomes with the top 5 measures.

In terms of data collection methods, the most commonly used is Observation (59 papers), followed by Self-report (53 papers), Instruments (29 papers), System log (20 papers), Inspection (8 papers) and Expert review (2 papers). Measureless outcomes (which have no data collection method) appear in 37 papers. If we use the outcomes as our unit of analysis, in Table 5.5, we can observe that the largest group are those that appear at least once with no measure (137), followed by Self-report (81), Observation (69), Instruments (35), System log (12), Inspection (2), and Expert review (2).

137 of the outcomes were presented without a measure linked to them on at least one occasion. Additionally, 90 outcomes never had a measure linked to them (i.e., were "exclusively measureless"). Satisfaction champions the list, being unmeasured 7 out of 15 times.

When we look at the outcomes that had no measure in any of the papers, “Immersion” is found at the first place. Analogously to Satisfaction, we expected Immersion

	Count	Unique outcomes	Unique measures	Top 5 outcomes	Top 5 measures	Top 5 outcome-measure pairs	Top 5 natures
All	81	227	649	Training outcomes (16), Satisfaction (15), Usefulness (12), Usability (9), Learning outcomes (9)	Time (24), Accuracy (21), Motion (13), R. Surgery time (10), R. Surgery accuracy (9)	Training outcomes, Accuracy (14), Training outcomes, Time (14), Training outcomes, Motion (12), Training outcomes, R. Surgery time (9), Training outcomes, R. Surgery accuracy (8)	O (59), S (53), I (29), L (20), IP (8)
Simulators	17	26	82	Training outcomes (16), Construct validity (8), Content validity (7), Face validity (7), Usefulness (5)	Accuracy (14), Time (14), Motion (12), R. Surgery time (10), R. Surgery accuracy (9)	Training outcomes, Accuracy (14), Training outcomes, Time (14), Training outcomes, Motion (12), Training outcomes, R. Surgery time (9), Training outcomes, R. Surgery accuracy (8)	O (16), L (14), S (12), I (2), IP (1)
Learning	13	74	90	Learning outcomes (9), Engagement (6), Enjoyment (5), Satisfaction (5), Motivation (5)	Time (4), Questionnaire of usefulness (2), Weight scale (1), Positive and negative affect schedule (1), Observation of verbal behaviors (1)	Usefulness, Questionnaire of usefulness (2), Learning outcomes, Time (2), Workload, Questionnaire of workload (1), Knowledge acquisition, R. Knowledge test on safety in construction (1), Enjoyment, Questionnaire of enjoyment (1)	O (10), S (9), I (4), L (2), E (1)
Psychology	12	33	142	Anxiety symptoms (3), Depression symptoms (3), Ptsd symptoms (3), Presence (2), Adjustment disorder symptoms (2)	Clinician-administered ptsd scale (3), Fear of flying scale (3), Questionnaire on attitudes toward flying (3), Behavioral avoidance test (3), Clinician global impression (3)	Anxiety symptoms, Social contexts inducing anxiety (2), Ptsd symptoms, Ptsd symptom scale (2), Anxiety symptoms, Liebowitz social anxiety scale (2), Anxiety symptoms, Personal report of confidence as a speaker (2), Ptsd symptoms, Clinician-administered ptsd scale (2)	S (10), O (9), I (6), L (1)
Post-stroke rehabilitation	11	26	155	Upper limb motor function (5), Balance (4), Activities of daily living (3), Motor function (2), Gait (2)	Berg balance scale (6), Timed up and go test (6), Fugl-meyer assessment (6), 6 minute walk test (5), Box and block test (5)	Upper limb motor function, Fugl-meyer assessment (4), Balance, Timed up and go test (3), Upper limb motor function, Modified ashworth scale (3), Activities of daily living, Modified barthel index (3), Balance, Berg balance scale (3)	O (9), S (7), I (6)
Cognition	8	41	106	Memory (4), Executive function (3), Attention (2), Balance (2), Mood (2)	fMRI (3), Trail making test (2), Automated neuropsychological assessment metrics (2), Delis-kaplan executive function system (2), EEG (2)	Memory, The rivermead behavioural memory test (2), Executive function, Wisconsin card sorting test (2), Attention, Paced auditory serial addition task (2), Wayfinding performance, Time (1), Cognitive function, Tower of london test (1)	O (6), S (4), I (4), L (1)
Surgery	6	16	31	System performance (6), Usefulness (3), Surgery performance (2), Surgery outcomes (2), Workload (1)	Accuracy (3), Tracking accuracy (3), Time (3), Registration accuracy (2), Segmentation accuracy (2)	System performance, Tracking accuracy (3), System performance, Registration accuracy (2), System performance, Rendering accuracy (2), Surgery performance, Accuracy (2), Surgery performance, Time (2)	IP (6), O (3), S (1), L (1)
Pain relief	5	18	29	Pain intensity (3), Pain unpleasantness (2), Anxiety symptoms (2), Pain (2), Time spent thinking about pain (2)	VAS of pain (3), Heart rate (2), Wong-Baker faces (1), Numeric pain rating scale (1), Blood oxygen saturation (1)	Pain, VAS of pain (2), Time spent thinking about pain, Reported time spent thinking about pain (1), Pain, Neuropathic pain symptom inventory (1), Anxiety symptoms, Burns specific pain anxiety scale (1), Anxiety symptoms, Children trauma screening questionnaire (1)	S (4), I (2), O (1)
Physical prevention	5	18	30	Balance (2), Gait (2), Tension (1), Calmness (1), Compliance (1)	Berg balance scale (2), Biodex Balance System (2), Force plate (2), Timed up and go test (1), The activation-deactivation adjective check list (1)	Balance, Berg balance scale (2), Dynamic balance, Force plate (1), Gait, Short physical performance battery (1), Gait, GAITRite system (1), Gait, GAITRite electronic walkway system (1)	S (4), O (3), I (3)
Multiple areas	3	35	23	User preference (1), Distracted attention (1), Intention to use (1), Immersion (1), Illusion (1)	EEG (2), Virtual Reality Symptom Questionnaire (1), Physiological data (1), Blink rate (1), EMG (1)	User preference, Subjective feedback (1), Cybersickness, Physiological data (1), Assembly task performance, Time (1), Block manipulation task performance, Accuracy (1), Block manipulation task performance, Time (1)	S (2), I (2), O (1), L (1), IP (1)
Industry	1	3	4	User acceptance (1), Satisfaction (1), Performance (1)	Informal user test of the system (1), Formal scientific setup user test of user acceptance (1), Formal scientific setup user test of performance (1), Expert review of the system (1)	User acceptance, Formal scientific setup user test of user acceptance (1), Satisfaction, Informal user test of the system (1), Satisfaction, Expert review of the system (1), Performance, Formal scientific setup user test of performance (1)	O (1), E (1)

to appear in more papers given its generality and relevance to XR. However, it appears in only three survey papers, none of which records what measure was used to weigh it. Among all the categories, Learning is the one with the most measureless and exclusively measureless outcomes.

Figure 5.1 shows an overview of the different outcomes, measures and their natures. The outcomes and measures are presented as nodes of the graph, distinguished by color (red and blue, respectively), and their size represents the amount of papers in which they appear. Outcome nodes are linked to the measures that are used to evaluate them, and the color of the link represents the nature of the measure for that particular pairing. In this network visualization we can see three large, distinct clusters: these are formed by outcomes that share measures among them, such as "Anxiety symptoms", "Depression symptoms" and "PTSD symptoms", which form a cluster of mainly self-report measures. The most prominent nodes in the graph are the measures of "Time", "Accuracy" and the outcome of "Training outcomes", which form a tightly knit cluster with observation and system log measures. A notable sub-graph is around the outcome of "System performance", which hoards several inspection measures. Finally, we can see an outer ring of measureless outcomes, not linked to any other node in the graph.

We also analyze the papers in “evaluation space”. Figure 5.2 is a scatter plot showing all 81 papers, colored according to their category. We used multidimensional scaling (MDS) to calculate the position of each paper according to it’s dissimilarity to all the others. To calculate the dissimilarity between the papers, we used a “bag-of-words” approach as in (MUNZNER, 2015). We encoded a boolean vector for each paper, with one bit for each distinct outcome, measure, nature, type of comparison, experiment design and type of evaluation, totalling 919 dimensions. For the dimensions that appeared in the paper, we assigned a value of one in the corresponding position along the vector; for the ones that didn’t, a zero. Then, we calculated the similarity between all pairs of vectors as: the number of dimensions minus the Hamming distance between the vectors, times two, divided by the sum of all true bits in both vectors; finally, we subtracted the similarity from one to obtain the dissimilarity, which served as input for the MDS. In effect, papers that are closer together in the graph share more in common in terms of outcomes, measures, natures, types of comparison, experiment designs and types of evaluation. Additionally, the size of each point in the scatter plot represents the amount of true bits – in other words, the diversity of evaluation information the paper presents. For example, if a paper presents several outcomes and measures, it will appear larger than a paper that shows only

Figure 5.1: Network visualization of outcomes, natures and measures. Outcomes and measures are represented as nodes, and their size is relative to the amount of papers they appear on. Nodes that appear in more than five papers are titled. The links between nodes are colored accordingly to the measure's nature; in cases where the same measure appears with two natures, the link is "striped" according to the proportion between natures. Outcomes or portion of outcomes that have no measures are colored lighter.

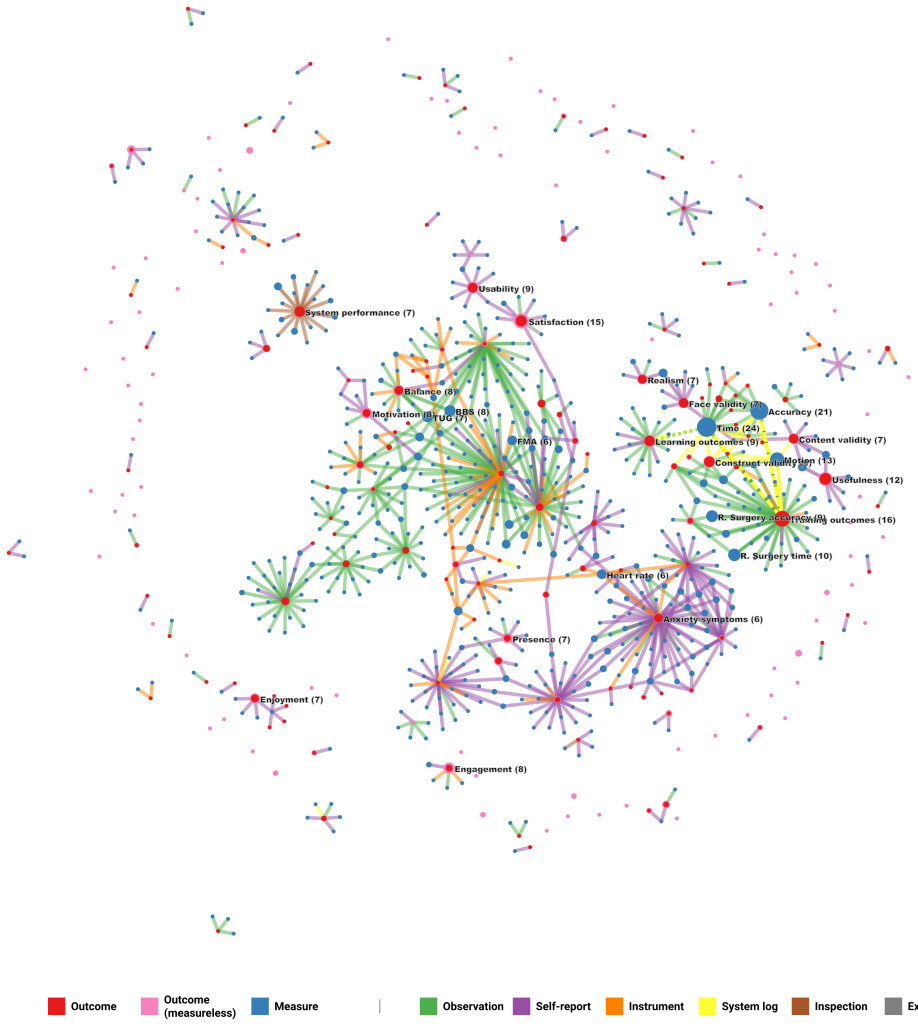


Figure 5.2: Scatter plot of the included surveys in evaluation space. The distance between each paper is relative to their similarity, in terms of the information on evaluation they contain. The size of each paper in the plot is relative to the amount of evaluation information it presents.

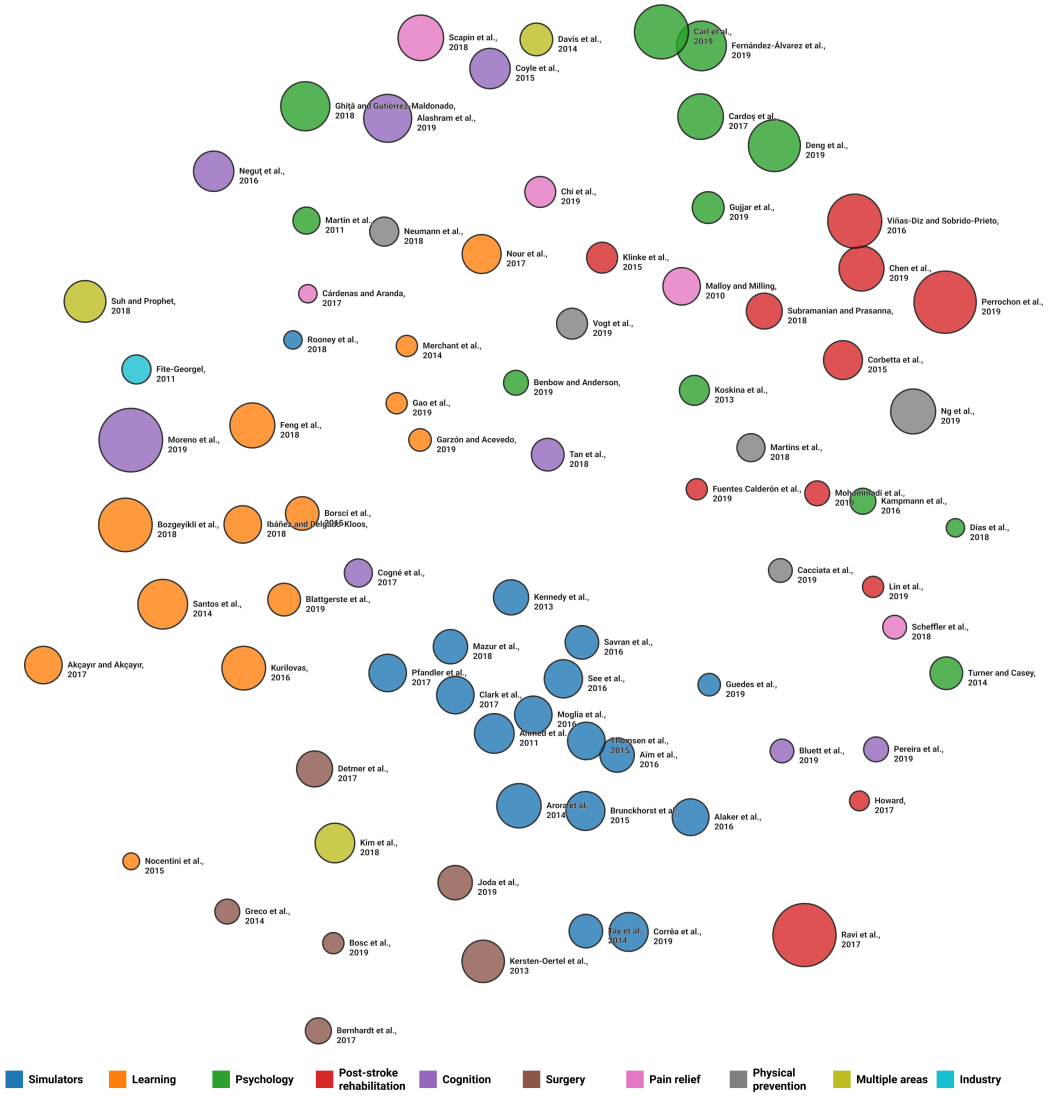
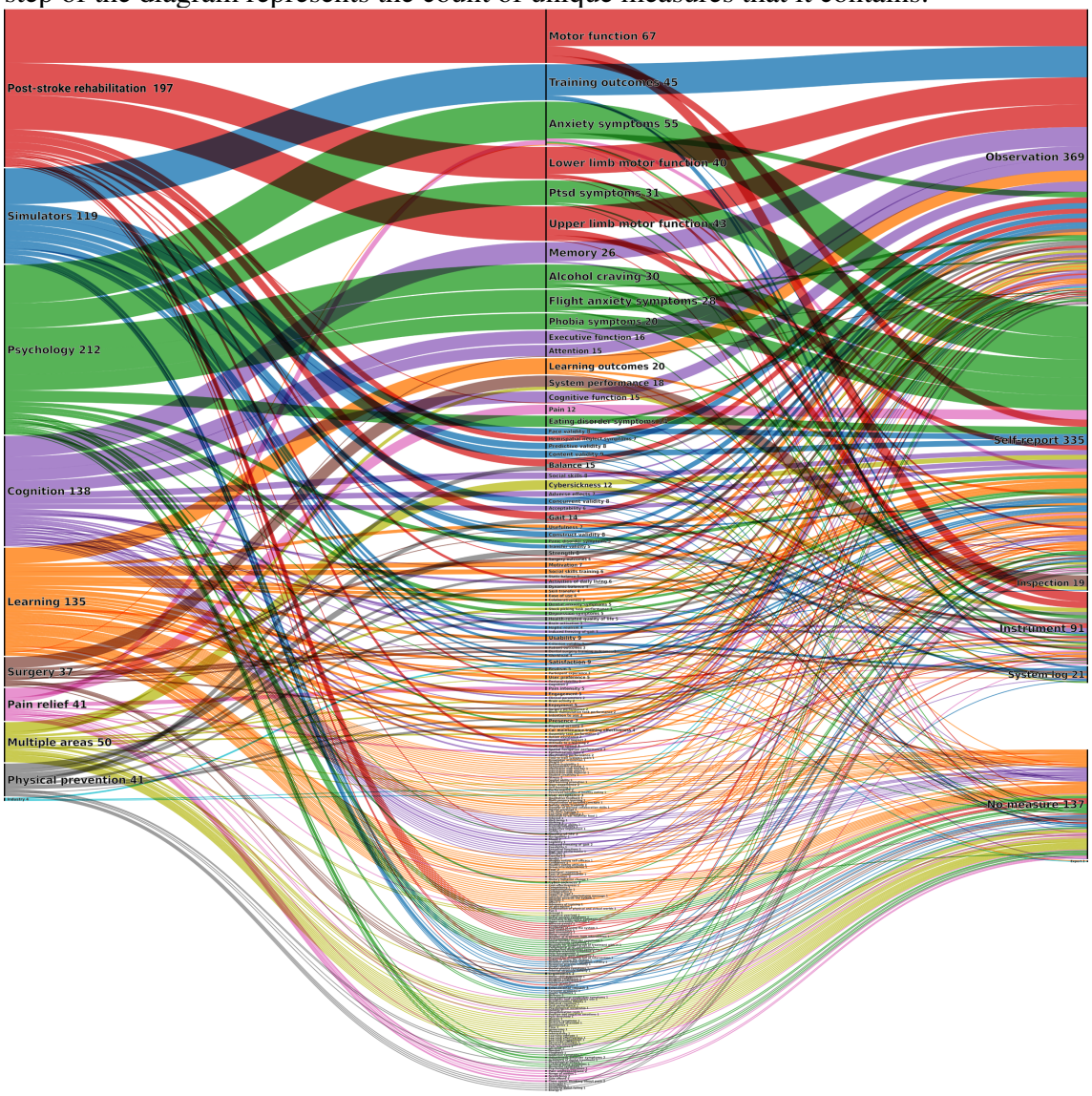


Figure 5.3: Sankey diagram of the categories, outcomes and natures. The number on each step of the diagram represents the count of unique measures that it contains.



a single outcome and measure.

Some categories, such as “Simulators” and “Learning” form clusters in the chart, suggesting that they share evaluation aspects to other papers in the category, while not sharing them with papers from other categories. “Psychology” and “Cognition” are intertwined in a large cluster in the chart, while being relatively large in area, hinting to a greater diversity of evaluation information in those categories. “Surgery” appears as two clusters, with one of those being comprised of smaller points, isolated from the rest of the papers, indicating that they carry a small diversity of evaluation information that is also unique to their category.

Finally, Figure 5.3 presents a Sankey diagram, showing all categories on the left

Table 5.6: Types of evaluation in the included papers

	All	Simulators	Learning	Psychology	Post-stroke rehabilitation	Cognition	Surgery	Pain relief	Physical prevention	Multiple areas	Industry
Count	81	17	13	12	11	8	6	5	5	3	1
Summative	77	17	11	12	11	8	4	5	5	3	1
Formative	14	1	7	-	-	1	2	-	1	1	1
Not described	1	-	1	-	-	-	-	-	-	-	-

side, and what outcomes and nature of measurement appear in the papers of each category. We can note that most outcomes (represented in the middle of the chart) are category-specific. The width of the flowline represents the number of measures it contains. We can also see that certain categories (such as Simulators and Post-Stroke Rehabilitation) have a handful of outcomes that contribute to the majority of evaluations, while others (such as Cognition and Learning) are scattered across many outcomes. We can also see that Psychology relies heavily on Self-report as a nature of measurement, while Post-Stroke Rehabilitation leans heavily towards Observation; the other categories are more balanced across the different measurement types. Finally, we can also see that some types of measurement are dominated by single categories, for example: "System log" has mostly entries from papers in the Simulator category, and "Inspection" mostly from Surgery. We added a "No measure" type for the purposes of this graph, and it receives flow from several categories, most markedly from Learning.

5.3.2 Types of evaluation

Table 5.6 provides an overview of the survey papers depending on the type of evaluation they report for the primary studies. Most papers included summative evaluations (n = 77), meaning that the focus of the evaluation was to find out the effects of the application on a certain outcome. This is in contrast to the smaller amount of studies that included formative evaluations (n = 14), which aimed to use the output of the evaluation explicitly to iterate over the application's design. In this latter category, we included evaluations that explicitly stated to be following a human-centered design process, as well as technical feasibility studies. One paper did not provide enough information to categorize its primary studies as summative or formative (AKÇAYIR; AKÇAYIR, 2017). Only 3

Table 5.7: Types of experiment design in the included papers

	All	Simulators	Learning	Psychology	Post-stroke rehabilitation	Cognition	Surgery	Pain relief	Physical prevention	Multiple areas	Industry
Count	81	17	13	12	11	8	6	5	5	3	1
Rct	44	11	1	10	10	4	1	4	3	-	-
Single group	27	11	7	2	-	2	2	2	1	-	-
Between-subjects	23	5	8	1	1	2	3	2	1	-	-
Non-equivalent	17	9	3	1	-	2	1	-	1	-	-
Case study	15	2	-	2	4	3	1	2	1	-	-
Within-subjects	14	-	3	1	4	2	-	-	2	2	-
Rct crossover	7	1	-	-	3	1	-	2	-	-	-
User study	5	-	3	-	-	-	-	-	-	1	1
Technical feasibility study	5	-	-	-	-	-	5	-	-	-	-
Not described	5	2	1	-	-	1	1	-	-	-	-
Formal scientific study	1	-	-	-	-	-	-	-	-	-	1
Focus group	1	-	1	-	-	-	-	-	-	-	-
Expert review	1	-	-	-	-	-	-	-	-	-	1
Experiments	1	-	-	-	-	-	-	-	-	1	-
Between subjects	1	-	-	-	-	-	-	-	1	-	-

studies included exclusively formative evaluations (BERNHARDT et al., 2017; NOCENTINI; ZAMBUTO; MENESINI, 2015; GRECO et al., 2014). For Bernhardt et al. (2017) and Greco et al. (2014), both in the “Surgery” category, the evaluations were mainly concerned with the technical feasibility of the applications, e.g., the tracking accuracy of an AR system. Lastly. Nocentini, Zambuto and Menesini (2015) reported only one XR primary study, which involved the design of a virtual environment collaboratively with the target users.

5.3.3 Study designs

The evaluations presented in the reviews adopted a range of different designs. We cataloged 13 different types of study designs, listed below from the most frequent to the less frequent. See Table 5.7 for the precise frequencies of the study designs for all papers and each category:

- Randomized controlled trial (RCT) is a design where at least two groups are subject to different interventions (e.g., one group received an XR-based intervention, and the other a non-XR-based intervention or a placebo). The groups are all equivalent (i.e., drawn from the same population at random), so that differences in the outcomes between groups can be more strongly linked to the intervention itself rather than differences between the groups' subjects.
- Single group is a design where only one group of subjects receives an intervention (and the outcome being studied is generally tested before and after the intervention);
- Between-subjects design employs two or more groups, each receiving a different intervention. It is the same nature as RCTs, but in a less robust manner because no randomization is performed.
- Non-equivalent design is a study employing at least two groups of subjects that are purposely not equivalent (e.g., a study involving a group of junior surgeons and a group of senior surgeons);
- Case studies are a more loosely defined design, where a single case or a series of cases are described;
- Crossover RCT is a similar design to RCT, but instead of each group receiving a different intervention, all groups receive all interventions, but in a different order. This means that each subject can act as its own control, but it also introduces possible issues if the order of intervention delivery might affect the results.
- Within-subjects design makes all subjects receive all interventions. Similar to Crossover RCTs, but not randomized.
- User study is also a more loosely defined design, involving the administration of an intervention to users without a formal experimental setting;
- Technical feasibility study is a design focusing on evaluating technical aspects of the XR application, and it does not necessarily involve users (e.g., gauging the tracking accuracy of an AR system for medicine in a controlled setting);

Table 5.8: Types of comparison in the included papers

	All	Simulators	Learning	Psychology	Post-stroke rehabilitation	Cognition	Surgery	Pain relief	Physical prevention	Multiple areas	Industry
Count	81	17	13	12	11	8	6	5	5	3	1
Conventional	50	13	9	5	7	6	2	2	5	1	-
No comparison	33	12	10	2	1	4	1	2	1	-	-
Passive	19	5	-	7	2	-	-	2	3	-	-
Not described	13	1	2	-	4	-	3	1	-	1	1
Non-equivalent	13	7	2	1	-	1	1	-	1	-	-
Vr to vr	12	1	3	1	-	2	-	2	2	1	-
Sub-intervention	12	-	-	4	5	-	-	2	1	-	-
Active	12	1	2	4	2	1	-	2	-	-	-
Placebo	8	-	-	4	2	-	-	2	-	-	-
Ar to ar	2	-	-	-	-	-	1	-	-	1	-
Xr to xr	1	-	1	-	-	-	-	-	-	-	-
Vr to ar	1	1	-	-	-	-	-	-	-	-	-

There were also other designs that appeared once each and did not provide enough information to place them in one of the types above: “Formal scientific study”, “Focus group”, “Expert review” and “Experiments”. Finally, 5 reviews did not provide any information on the study designs.

5.3.4 Comparison to XR

In the primary studies, XR applications are compared to several different interventions (or not compared to anything). We cataloged 11 types of comparisons, listed below from the most frequent to the least frequent one. See Table 5.8 for the precise frequencies of each comparison type for all papers and each category:

- Conventional: XR is compared to another intervention that is referred to as “conventional” or “traditional” in the domain, and does not involve XR;
- No comparison: XR is explicitly not compared to anything in the evaluation;
- Passive: XR is compared to not receiving any intervention (e.g., a “waitlist” condition);
- Non-equivalent: directly linked to the non-equivalent design, is mostly a compari-

son between groups of subjects, and not between interventions;

- VR to VR: when two or more VR interventions are compared.
- Sub-intervention: when XR is used as an auxiliary intervention, and is compared to the main intervention by itself;
- Active: XR is compared to another experimental intervention that is also not "conventional";
- Placebo: XR is compared to a placebo, such as attention control or an informational pamphlet;
- AR to AR: when two or more AR interventions are compared;
- XR to XR: when AR/VR/MR interventions are compared among themselves;
- VR to AR: when a VR and an AR intervention are compared;

Additionally, 33 studies did not report enough information on comparisons in the evaluations.

6 ANALYSIS AND DISCUSSION

In this section we report the findings for each category (Simulators, Learning, Psychology, Post-stroke Rehabilitation, Cognition, Surgery, Pain Relief, Physical Prevention, Multiple areas, and Industry), while answering our research questions.

6.1 Simulators

Seventeen studies focused on simulator training for medical education (GUEDES et al., 2019; CORRÊA et al., 2019; ROONEY et al., 2018; MAZUR et al., 2018; PFANDLER et al., 2017; CLARK et al., 2017; SEE et al., 2016; SAVRAN et al., 2016; MOGLIA et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; AÏM et al., 2016; THOMSEN et al., 2015; BRUNCKHORST et al., 2015; TAY; KHAJURIA; GUPTE, 2014; ARORA et al., 2014; KENNEDY; MALDONADO; COOK, 2013; AHMED et al., 2011). Most of them sought evidence on the effectiveness of training using simulators.

6.1.1 What is XR for this category of studies?

Among the 17 studies, only four presented some definition of XR. Corrêa et al. (2019) defined VR as 3D virtual environments that allow for real-time interaction, and AR also as real-time, but presenting both real and virtual elements that allow for 3D interaction. For these authors, VR and AR can be delivered with regular monitors, and not necessarily with HMDs or special glasses. Rooney et al. (2018) defines VR and haptic systems together, as systems that include a physical representation and sensors that inform the computer of the user's movements. Pfandler et al. (2017) defines VR as a potentially multisensory virtual environment. Instruments can be used to interact with the virtual environment, which can provide force feedback. MR is defined as the combination of VR and reality, and AR as the addition of virtual components (such as the optimal entry angle of a patient's back) to real environments. Alaker, Wynn and Arulampalam (2016) simply defined VR in the context of simulation, as computer software and hardware similar to that used in surgery.

The VR simulators in all 17 studies were typically described solely by their make and model, with some exceptions (CORRÊA et al., 2019; ROONEY et al., 2018; BRUNCK-

HORST et al., 2015; ARORA et al., 2014; KENNEDY; MALDONADO; COOK, 2013). In these works the simulators were described in terms of the system features (e.g., type of display, characteristics of the virtual environment, whether haptic feedback is employed). Some (CORRÊA et al., 2019; PFANDLER et al., 2017) also included AR and MR simulators, but without details.

6.1.2 What's the motivation for the use of XR?

The main reason for employing VR simulators in medical education is that it enables practicing surgeons to train without risking the safety of patients (CORRÊA et al., 2019; ROONEY et al., 2018; MAZUR et al., 2018; PFANDLER et al., 2017; CLARK et al., 2017; SEE et al., 2016; SAVRAN et al., 2016; MOGLIA et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; THOMSEN et al., 2015; BRUNCKHORST et al., 2015; TAY; KHAJURIA; GUPTE, 2014; ARORA et al., 2014; KENNEDY; MALDONADO; COOK, 2013; AHMED et al., 2011). In contrast to training with real patients, VR simulators allow for potentially limitless training opportunities (CORRÊA et al., 2019; MAZUR et al., 2018; PFANDLER et al., 2017; SEE et al., 2016; SAVRAN et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; AÏM et al., 2016; ARORA et al., 2014). A related advantage is the ability to increase the trainees' exposure to complex surgical scenarios, such as rare diseases (GUEDES et al., 2019; SEE et al., 2016; KENNEDY; MALDONADO; COOK, 2013).

Secondly, VR systems can record performance metrics automatically, such as the accuracy of instrument use and the number of areas incorrectly touched (CORRÊA et al., 2019; MAZUR et al., 2018; PFANDLER et al., 2017; CLARK et al., 2017; AÏM et al., 2016; THOMSEN et al., 2015; TAY; KHAJURIA; GUPTE, 2014; AHMED et al., 2011). These metrics can then be used to assess, compare and certify trainees objectively (CORRÊA et al., 2019; MAZUR et al., 2018; PFANDLER et al., 2017; AÏM et al., 2016; THOMSEN et al., 2015; TAY; KHAJURIA; GUPTE, 2014; AHMED et al., 2011).

Thirdly, VR can enhance learning through its ability to provide real-time feedback to trainees (MAZUR et al., 2018; SEE et al., 2016; SAVRAN et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; TAY; KHAJURIA; GUPTE, 2014). VR also requires less supervision by trained professionals (GUEDES et al., 2019; ALAKER; WYNN; ARULAMPALAM, 2016; AÏM et al., 2016). Finally, the virtual scenarios can be adapted to the trainee level, who can also choose to focus on a specific part of the surgical per-

formance (CORRÊA et al., 2019; MAZUR et al., 2018; ALAKER; WYNN; ARULAMPALAM, 2016).

Some authors (GUEDES et al., 2019; CORRÊA et al., 2019; MAZUR et al., 2018; PFANDLER et al., 2017; TAY; KHAJURIA; GUPTE, 2014) argue that VR is less costly than traditional training, which involves the costs of e.g., keeping an animal or cadaver laboratory. VR is also argued to be more realistic than these other methods (CORRÊA et al., 2019; PFANDLER et al., 2017; CLARK et al., 2017; AÏM et al., 2016). However, other authors (THOMSEN et al., 2015; ARORA et al., 2014) question whether realism is truly necessary for the training effectiveness of a simulator.

6.1.3 What did the evaluations on the primary studies focus on?

Most of the primary studies evaluated the effectiveness of simulators as training tools. This outcome was generally expressed in terms of time to complete tasks or other objective performance metrics captured by the simulator itself (such as length of the trajectory of the instruments, accuracy of instrument use, percentage of simulated tumor resected) (GUEDES et al., 2019; CORRÊA et al., 2019; PFANDLER et al., 2017; CLARK et al., 2017; SEE et al., 2016; SAVRAN et al., 2016; MOGLIA et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; AÏM et al., 2016; THOMSEN et al., 2015; BRUNCKHORST et al., 2015; TAY; KHAJURIA; GUPTE, 2014; ARORA et al., 2014; KENNEDY; MALDONADO; COOK, 2013; AHMED et al., 2011). Another way to measure the effectiveness of the simulator was through scales designed to evaluate performance in surgery, such as the Global Rating Scale (GRS), the Objective Structured Assessment of Technical Skills (OSATS), and the Global Operative Assessment of Laparoscopic Skills (GOALS) (GUEDES et al., 2019; CORRÊA et al., 2019; SEE et al., 2016; SAVRAN et al., 2016; MOGLIA et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; AÏM et al., 2016; THOMSEN et al., 2015; BRUNCKHORST et al., 2015; TAY; KHAJURIA; GUPTE, 2014; KENNEDY; MALDONADO; COOK, 2013; AHMED et al., 2011). Finally, patient outcomes (i.e., the clinical effects the surgery had on the patient) are used as a measure in (MAZUR et al., 2018; KENNEDY; MALDONADO; COOK, 2013).

The second main outcome of the evaluation of simulators is their validity (CORRÊA et al., 2019; MAZUR et al., 2018; PFANDLER et al., 2017; CLARK et al., 2017; SEE et al., 2016; SAVRAN et al., 2016; MOGLIA et al., 2016; AÏM et al., 2016; THOM-

SEN et al., 2015; BRUNCKHORST et al., 2015; TAY; KHAJURIA; GUPTE, 2014; ARORA et al., 2014; AHMED et al., 2011). The validity of a simulator signifies how well the performance of a task on the simulator is representative of the performance of an actual surgical task. It can be broken down into various types, such as face, content, construct, discriminant, concurrent, transfer and criterion validity. The categories of validity change depending on the validity framework adopted by each of the studies. However, some patterns emerge, such as a Likert scale rating of perceived realism being used as a measure of face validity (PFANDLER et al., 2017; SAVRAN et al., 2016; MOGLIA et al., 2016; BRUNCKHORST et al., 2015; ARORA et al., 2014; AHMED et al., 2011) and the difference of performance scores between junior and experts as construct or criterion validity (MAZUR et al., 2018; PFANDLER et al., 2017; CLARK et al., 2017; ARORA et al., 2014).

Satisfaction of the learners and teachers with the simulator was also included in some of the studies (ROONEY et al., 2018; MAZUR et al., 2018; PFANDLER et al., 2017; SAVRAN et al., 2016; BRUNCKHORST et al., 2015; ARORA et al., 2014; KENNEDY; MALDONADO; COOK, 2013; AHMED et al., 2011). This outcome was usually assessed via a Likert scale rating of perceived usefulness.

Other outcomes included were ease of use (MOGLIA et al., 2016; ARORA et al., 2014) and engagement (PFANDLER et al., 2017). (CLARK et al., 2017; ARORA et al., 2014; AHMED et al., 2011) included qualitative feedback from the participants. For example, in the survey by Cogné et al. (2017), one study notes that some users found it hard to execute the neurosurgical tasks in the simulator without being able to touch the patient's forehead to search for anatomical landmarks.

Three meta-analyses were included (GUEDES et al., 2019; ALAKER; WYNN; ARULAMPALAM, 2016; KENNEDY; MALDONADO; COOK, 2013). Guedes et al. (2019), compared the effectiveness of VR simulators and box-trainers (a common training tool for minimally invasive surgery) for learning technical skills. The outcomes were the time to complete training tasks (where VR fared better in some cases) and performance score and time to complete a minimally invasive surgery (where there was no difference). The authors argue that speed is not a good measure of quality in surgery. Alaker, Wynn and Arulampalam (2016) also compares VR simulators to box-trainers and to video trainers (which is a version of the box-trainer equipped with sensors). VR simulators were found better than video trainers or no training at all in time to perform tasks and, in two studies that evaluated live operative performance, also led to higher GRS scores

than box-trainers. However, the authors argue none of these outcomes reflects patient outcomes directly. In a subgroup analysis of haptic and non-haptic VR simulators versus box-trainers, the authors found that only haptic-capable VR was equivalent or superior to box-trainers. The authors argue that since VR simulators are much more costly than either of the other options, further cost-benefit analysis is necessary. Finally, Kennedy, Maldonado and Cook (2013) analyzed VR and non-VR simulators for the outcomes of perceived usefulness and satisfaction, where non-VR simulators fared better. They also included studies comparing VR simulators to no training at all, where VR simulators had a positive effect. All three meta-analyses reported significant heterogeneity among the included VR studies.

6.1.4 What are the criticisms to the primary studies?

Regarding outcomes, the main criticism is the lack of a clear link between improved performance in the simulator and improved patient outcomes (MAZUR et al., 2018; PFANDLER et al., 2017; CLARK et al., 2017; SEE et al., 2016; MOGLIA et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; AÏM et al., 2016; THOMSEN et al., 2015) (i.e., transfer validity). For example, Mazur et al. (2018) argues that without the inclusion of such outcomes, the increased performance after training sessions in the simulator might be a measure of how capable the trainees are in using the simulator, rather than a measure of surgical skill. Other criticisms regarding outcomes are the lack of a standardized proficiency parameter (e.g., to differentiate junior and expert trainees) (GUEDES et al., 2019; CORRÊA et al., 2019; MAZUR et al., 2018; ARORA et al., 2014; AHMED et al., 2011), and the reliance on subjective measures such as the user's opinion of their own performance improvement, or ratings of realism (SEE et al., 2016; SAVRAN et al., 2016; ARORA et al., 2014).

Another criticism is of the methodological quality of the primary studies, mostly due to the lack of randomized controlled studies (SEE et al., 2016; MOGLIA et al., 2016; KENNEDY; MALDONADO; COOK, 2013), lack of comparison to traditional training (MAZUR et al., 2018; CLARK et al., 2017; SEE et al., 2016), possible biases (PFANDLER et al., 2017; SEE et al., 2016; ARORA et al., 2014; KENNEDY; MALDONADO; COOK, 2013) and small sample sizes (CLARK et al., 2017; SEE et al., 2016; AÏM et al., 2016).

6.1.5 What paths for future research are suggested?

The main future path for research proposed in the Simulators category studies is to justify the costs of VR simulators (CLARK et al., 2017; SEE et al., 2016; MOGLIA et al., 2016; ALAKER; WYNN; ARULAMPALAM, 2016; THOMSEN et al., 2015; BRUNCKHORST et al., 2015; TAY; KHAJURIA; GUPTE, 2014).

Another future path is trying to understand what makes the intervention effective (CORRÊA et al., 2019; SAVRAN et al., 2016; KENNEDY; MALDONADO; COOK, 2013) by comparing different interventions and instructional designs, for example. Some authors (ROONEY et al., 2018; BRUNCKHORST et al., 2015) argue that non-technical skills (i.e., social skills in the operating room) should also be evaluated. Rooney et al. (2018) additionally proposes that simulators should also target non-physician members of the team.

6.2 Learning

Thirteen studies focused on learning using XR. Six (SANTOS et al., 2014; KURILOVAS, 2016; GARZÓN; ACEVEDO, 2019; IBÁÑEZ; DELGADO-KLOOS, 2018; MERCHANT et al., 2014; AKÇAYIR; AKÇAYIR, 2017) were concerned with education, from preschool to postgraduate levels. Two focused on learning for cognitively impaired people (BOZGEYIKLI et al., 2018; BLATTGERSTE; RENNER; PFEIFFER, 2019). Three focused on training specific subjects: health and safety in construction (GAO; GONZALEZ; YIU, 2019), evacuation (FENG et al., 2018), and car maintenance (BORSCI; LAWSON; BROOME, 2015). The remaining two works studied technological interventions for nutrition (NOUR et al., 2017) and cyber-bullying (NOCENTINI; ZAMBUTO; MENESINI, 2015), and involved learning about these subjects as an outcome.

6.2.1 What is XR for this category of studies?

Five studies focused on AR (SANTOS et al., 2014; BLATTGERSTE; RENNER; PFEIFFER, 2019; GARZÓN; ACEVEDO, 2019; IBÁÑEZ; DELGADO-KLOOS, 2018; AKÇAYIR; AKÇAYIR, 2017), of which only one (BLATTGERSTE; RENNER; PFEIFFER, 2019) did not provide an explicit definition. All other four define AR as a technol-

ogy that allows the integration of virtual objects with the real environment. For Santos et al. (2014), Ibáñez and Delgado-Kloos (2018) this integration is, by definition, done in real-time, and "anchored" to the real environment through registration. However, (SANTOS et al., 2014) admits some relaxations to this definition in the included studies, such as overlaying 2D images rather than 3D objects, and imperfect registration. For (GARZÓN; ACEVEDO, 2019), virtual objects can be overlaid on the real environment through senses other than sight (sound, for example), and thus advocates for the definition by Akçayır and Akçayır (2017) rather than the narrower definition of AR as augmenting the visual field found in (CAUDELL; MIZELL, 1992; AZUMA, 1997). Specifically, the included primary studies involved HMDs, projectors, desktop monitors and handheld devices (SANTOS et al., 2014; BLATTGERSTE; RENNER; PFEIFFER, 2019; AKÇAYIR; AKÇAYIR, 2017).

Five studies focused on VR (BOZGEYIKLI et al., 2018; FENG et al., 2018; NOUR et al., 2017; NOCENTINI; ZAMBUTO; MENESINI, 2015; MERCHANT et al., 2014). Three studies (NOUR et al., 2017; NOCENTINI; ZAMBUTO; MENESINI, 2015; MERCHANT et al., 2014) did not provide a definition of VR. In two studies (BOZGEYIKLI et al., 2018; FENG et al., 2018), VR aims at immersing users in a virtual environment. This immersion can vary according to the technologies employed (FENG et al., 2018), and high levels of both immersion and interaction are what distinguish VR from other computer technologies (BOZGEYIKLI et al., 2018). In practice, the included VR applications were based on immersive (HMD or CAVE) (BOZGEYIKLI et al., 2018; FENG et al., 2018) and non-immersive (desktop-based) devices (BOZGEYIKLI et al., 2018; NOUR et al., 2017; NOCENTINI; ZAMBUTO; MENESINI, 2015; MERCHANT et al., 2014).

Other three studies encompassed AR, MR and VR (KURILOVAS, 2016; GAO; GONZALEZ; YIU, 2019; BORSCI; LAWSON; BROOME, 2015). Kurilovas (2016) did not provide a definition. For Gao, Gonzalez and Yiu (2019), VR is a computer technology – that can make use of HMDs or projection-based displays – to create realistic virtual environments that elicit a feeling of physically existing in it. AR and MR are defined as "cutting-edge visualization technologies" that superimpose visual content on the real world. Also, they define MR as an "evolution of AR" because it anchors virtual objects to the real world and allows interaction. This specifically contradicts the definition by Borsci, Lawson and Broome (2015), that considers MR as a continuum that spans from reality to virtuality, encompassing augmented reality and augmented virtuality

(See (MILGRAM; KISHINO, 1994)). The reality-virtuality continuum is also acknowledged by (SANTOS et al., 2014).

6.2.2 What's the motivation for the use of XR?

The main reason for the use of AR was increased motivation, which could lead to better learning outcomes (BLATTGERSTE; RENNER; PFEIFFER, 2019; GARZÓN; ACEVEDO, 2019; IBÁÑEZ; DELGADO-KLOOS, 2018). For Blattgerste, Renner and Pfeiffer (2019), one of the drivers of such increased motivation can be the gamification aspects of the AR systems employed. Motivation stems from the characteristics of AR media: sensory immersion, navigation and manipulation (IBÁÑEZ; DELGADO-KLOOS, 2018).

For Santos et al. (2014), the AR affordances of real world annotation, contextual visualization and vision-haptic visualization can be leveraged to design better learning experiences. The affordance of contextualized information is especially important for cognitively impaired people (BLATTGERSTE; RENNER; PFEIFFER, 2019). On the other hand, Garzón and Acevedo (2019) argues that there might be a negative effect of information overload on young learners using AR, while for other author (IBÁÑEZ; DELGADO-KLOOS, 2018), there is no evidence on AR decreasing cognitive load or enhancing spatial ability.

Motivation is also the main reason for the use of VR (FENG et al., 2018; NOUR et al., 2017; NOCENTINI; ZAMBUTO; MENESINI, 2015; MERCHANT et al., 2014). Especially the game elements are linked to engagement and motivation, which can lead to enhanced learning (MERCHANT et al., 2014). The authors state that games should provide autonomy, identity and interactivity. Novelty effects of VR are also acknowledged as a driver for motivation. (FENG et al., 2018) argues immersion leads to full engagement and high emotional and physiological arousal.

Personalisation (KURILOVAS, 2016; BORSCI; LAWSON; BROOME, 2015) and motivation (GAO; GONZALEZ; YIU, 2019; BORSCI; LAWSON; BROOME, 2015) are the main arguments for the use of XR systems for learning. Additionally, (GAO; GONZALEZ; YIU, 2019) states that XR systems can rely less on textual information, and thus can be useful for illiterate users. Presence, flow and identification with the virtual characters are cited as drivers for motivation.

6.2.3 What did the evaluations on the primary studies focus on?

All studies investigated the effectiveness of XR systems on learning outcomes. These were mainly assessed using tests for measuring student performance.

Several studies also included usability as an outcome (BOZGEYIKLI et al., 2018; SANTOS et al., 2014; BLATTGERSTE; RENNER; PFEIFFER, 2019; KURILOVAS, 2016; IBÁÑEZ; DELGADO-KLOOS, 2018; FENG et al., 2018; AKÇAYIR; AKÇAYIR, 2017; BORSCI; LAWSON; BROOME, 2015). Santos et al. (2014) argues that the included primary studies did not really measure usability (or the related constructs of ease of use, usefulness and intention to use) per se, but rather perceived usability. They note that the direct evaluation of ease of use through time spent on tasks and number of errors were seldom carried out in the primary studies. Some studies surveyed by Blattgerste, Renner and Pfeiffer (2019) included both perceived usability measures, tasks' completion time and errors. They found that cognitively impaired people using AR action assistance systems made more errors, which would indicate a need to revise the design principles of AR assistance systems.

The outcome of motivation is included in studies surveyed by several authors (SANTOS et al., 2014; BLATTGERSTE; RENNER; PFEIFFER, 2019; KURILOVAS, 2016; IBÁÑEZ; DELGADO-KLOOS, 2018; FENG et al., 2018; AKÇAYIR; AKÇAYIR, 2017; BORSCI; LAWSON; BROOME, 2015). These are mostly measured using Self-report scales. The outcome of immersion is included in two surveys (KURILOVAS, 2016; IBÁÑEZ; DELGADO-KLOOS, 2018), but no measures are provided.

Three of the studies were meta-analyses (SANTOS et al., 2014; GARZÓN; ACEVEDO, 2019; MERCHANT et al., 2014). Santos et al. (2014), in addition to their systematic review, performed a meta-analysis of seven studies that reported effect sizes of AR on learning. They found a moderate effect size of AR on learning, but note that the implementation of the interventions and the control conditions varied greatly. The second survey (GARZÓN; ACEVEDO, 2019) surveyed 64 studies and a moderate effect of AR on learning gains was found. Type of control, learning environment (formal, informal, both), level of education and field of education were tested as possible moderators. Informal learning environments and higher education levels were found to positively moderate the learning gains. The area of education also moderated the effects (with Engineering and Arts showing higher effects). The authors note that the findings of the moderator analysis can be biased by the small number of studies found in some of the categories.

Finally, Merchant et al. (2014) investigated the effectiveness of desktop-based VR instruction on students' learning outcomes. In a comparison of simulations, games and virtual worlds, the authors found that games had the greatest effects on learning outcomes, but that these outcomes diminished with more time in-game, what the authors pointed as evidence of the novelty effect of the technology waning off. In addition to exposure time, other moderators explored were type of learning outcome (e.g., knowledge or skill) and type of instructional feedback.

6.2.4 What are the criticisms to the primary studies?

The main criticism of the primary studies were the lack of longitudinal designs (BLATTGERSTE; RENNER; PFEIFFER, 2019; GARZÓN; ACEVEDO, 2019; GAO; GONZALEZ; YIU, 2019; IBÁÑEZ; DELGADO-KLOOS, 2018; NOUR et al., 2017; AKÇAYIR; AKÇAYIR, 2017; BORSCI; LAWSON; BROOME, 2015) and small sample sizes (BOZGEYIKLI et al., 2018; BLATTGERSTE; RENNER; PFEIFFER, 2019; KURILOVAS, 2016; GAO; GONZALEZ; YIU, 2019).

Regarding the evaluations performed in the primary studies, Santos et al. (2014) argue that the comparison to traditional learning methods were not fair. The usage of ad-hoc questionnaires as outcome measures is criticized by them and others (IBÁÑEZ; DELGADO-KLOOS, 2018; BORSCI; LAWSON; BROOME, 2015). For Borsci, Lawson and Broome (2015), the studies should have used the already available and validated measures of usability and presence, as well as report on the incidence of cybersickness.

6.2.5 What are the paths for future research for the authors in this area?

Future research should aim to discover which specific properties of VR (e.g., output modality, visual fidelity, characteristics of the VE, usage of virtual avatars) have effects on learning outcomes (BOZGEYIKLI et al., 2018). Similarly for AR (GARZÓN; ACEVEDO, 2019), future research should take into account the type of augmentation, as well as the participants characteristics such as motivation, spatial competence and attitude towards technology as possible moderators of the effectiveness of the systems on learning. These authors also suggest that future research should try to cover the gaps in age and field of education they encountered. The further exploration on how student's characteris-

tics might affect or moderate outcomes is echoed by other authors (KURILOVAS, 2016; IBÁÑEZ; DELGADO-KLOOS, 2018; BORSCI; LAWSON; BROOME, 2015). Santos et al. (2014) argue that future research should evaluate usability quantitatively (e.g., using time on task and errors made) rather than only it is perceived. For Ibáñez and Delgado-Kloos (2018), the design of the AR interventions should be improved, so that all AR affordances are usable by students and teachers. In other survey (AKÇAYIR; AKÇAYIR, 2017), the most common challenge for the use of AR in educational settings was that "AR was difficult for students to use". The authors argue that lack of usability can impact learning effectiveness, and cause cognitive overload. However, they note that ease of use and reduced cognitive load are cited as advantages of AR in other studies. Thus, they suggest future research should try to clarify these conflicting conclusions – finding if there is in fact a usability issue, and in case there is, where does it come from (poor interface design, technical problems, or the teacher's inexperience with technology or negative attitude towards it). The authors also suggest AR applications design should take the students opinions and preferences into account.

In terms of outcomes that should be considered by future research, Akçayır and Akçayır (2017) suggests studies should take not only academic achievement into account, but learners satisfaction and confidence as well. For others (GAO; GONZALEZ; YIU, 2019), effectiveness of training should be measured in the future in terms of injury rate reduction, behavior alteration and tests of retention of knowledge.

According to the reviewed studies, the design of the applications on the primary studies could be improved in two main ways: the adoption of a clear underlying pedagogic framework (KURILOVAS, 2016; GARZÓN; ACEVEDO, 2019; IBÁÑEZ; DELGADO-KLOOS, 2018; MERCHANT et al., 2014), and by finding and making use of empirically proven design principles (BOZGEYIKLI et al., 2018; AKÇAYIR; AKÇAYIR, 2017).

6.3 Psychology

Twelve studies reviewed the use of VR as a psychological intervention. Most (8) focused on Virtual Reality Exposure Therapy (VRET) (GUJJAR et al., 2019; FERNÁNDEZ-ÁLVAREZ et al., 2019; DENG et al., 2019; CARL et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; KOSKINA; CAMPBELL; SCHMIDT, 2013). One survey (TURNER; CASEY, 2014) encompassed both VRET as well as other types of VR therapy, while other (GHIŢĂ;

GUTIÉRREZ-MALDONADO, 2018) focused on Cue Exposure Therapy (CET). The two remaining studies, on gamification in depression care (DIAS; BARBOSA; VIANNA, 2018), and networked communications (MARTIN et al., 2011) each included one primary study using VR.

6.3.1 What is XR for this category of studies?

Eight studies did not explicitly define VR. The other five reviews (DENG et al., 2019; GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; TURNER; CASEY, 2014) defined VR as computer-generated simulations. For three surveys (DENG et al., 2019; GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018; CARDOŞ; DAVID; DAVID, 2017), these simulations allow for the systematic exposure of the patient to feared stimuli, one of them (DENG et al., 2019) using an HMD. Only three studies (KAMPMANN; EMMELKAMP; MORINA, 2016; TURNER; CASEY, 2014; KOSKINA; CAMPBELL; SCHMIDT, 2013) (including two among the eight on VRET) provided descriptions of the VR applications included in terms of the virtual scenarios and their purposes: for example, virtual social situations or other feared stimuli are presented and interacted with by the participant in a realistic, highly controllable setting, that might include multisensory stimuli (e.g., auditory, olfactory, in addition to visual) (KAMPMANN; EMMELKAMP; MORINA, 2016; KOSKINA; CAMPBELL; SCHMIDT, 2013). For cue exposure therapy (in alcohol misuse), both projectors, stereoscopic monitors and HMDs were included, as well as VEs with different settings (e.g., a japanese pub, hotel bars, whether avatars were used to induce peer-pressure) (GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018). While the descriptions of what actually constituted VR in the VRET reviews, especially in terms of hardware, are lacking, this might be due to a somewhat long tradition of VRET, to the point where *in virtuo* is used to describe exposure to virtual stimuli (alongside *in vivo* and *in imago*) (DENG et al., 2019).

6.3.2 What's the motivation for the use of XR?

The main motivations for the use of VR in psychological interventions are its personalization, higher tolerability compared to *in vivo* exposure, and potential cost-

effectiveness.

The ability to personalize and control the VR environment is an advantage (FERNÁNDEZ-ÁLVAREZ et al., 2019; DENG et al., 2019; GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; KOSKINA; CAMPBELL; SCHMIDT, 2013). However, Koskina, Campbell and Schmidt (2013) believes changing the exposure protocol in the middle of a session is less achievable in VR than in conventional therapy. That might be an usability limitation of the application employed in the primary studies (failing to enable the therapist to modify the environment during the session).

In VRET, VR exposure is considered more tolerable than *in vivo* exposure because the patients perceive the virtual exposure as safer (FERNÁNDEZ-ÁLVAREZ et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; KOSKINA; CAMPBELL; SCHMIDT, 2013); or less stigmatized (CARL et al., 2019). Carl et al. (2019) also argue VRET can overcome mobility and geographic limitations – although that adds a remote component to the therapy.

Cost-effectiveness is reported as a motivation (GUJJAR et al., 2019; DENG et al., 2019; CARL et al., 2019; TURNER; CASEY, 2014) with Carl et al. (2019) arguing that VRET has much lower costs than traditional psychotherapy.

Especially, some authors (GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018) note that the possibility of creating life-like simulation is useful for CET, to increase the transferability of the treatment effects to the real world.

6.3.3 What did the evaluations on the primary studies focus on?

All evaluations sought evidence of the effectiveness of using virtual reality in treating various disorders – in other words, gauging the impact of the use of VR on the participants' symptoms.

Thus, the evaluations focused mostly on clinical outcomes (anxiety, post-traumatic stress disorder, depression), by means of domain-specific measures. These measures were either self-reported (such as the Liebowitz social anxiety scale), observation-based (such as the behavioral avoidance test or – in the case of dental anxiety – the avoidance of seeking dental treatment after the intervention), or instrumented (heart rate, skin conductance).

Two studies also reported measures of cybersickness (GUJJAR et al., 2019; BENBOW; ANDERSON, 2019). Benbow and Anderson (2019) specifically studied attrition

(drop-out) from VRET. They collected the number of drop-outs and the reasons for doing so and found that the most common reasons were failure to immerse in the VR environment, cybersickness, not-normal vision (i.e., myopia) and discomfort in communicating with a therapist without seeing them. In contrast, the most common reason for dropping out of *in vivo* exposure was fear of exposure. The authors suggest that, were the primary studies not randomized, attrition in VR could be further reduced, since participants with fear of exposure to the real stimuli could choose VRET instead.

Seven of the studies were meta-analyses (FERNÁNDEZ-ÁLVAREZ et al., 2019; DENG et al., 2019; CARL et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; TURNER; CASEY, 2014), what contributes for the focus on effectiveness. Apart from the aforementioned study (BENBOW; ANDERSON, 2019) that focused on attrition, another meta-analysis focused on deterioration rates (FERNÁNDEZ-ÁLVAREZ et al., 2019) during VRET, and found that VR does not seem to cause more adverse effects than conventional methods. The other meta-analyses focused on comparing VRET effectiveness against in vivo exposure and control conditions, which show that VRET is equivalent (CARL et al., 2019) or slightly superior to *in vivo* exposure (CARDOŞ; DAVID; DAVID, 2017), and equivalent to other types of therapy (DENG et al., 2019; KAMPMANN; EMMELKAMP; MORINA, 2016). This general equivalence of VRET to conventional exposure is deemed due to both relying on the same mechanism of habituation (CARDOŞ; DAVID; DAVID, 2017). Finally, Turner and Casey (2014) studied several types of VR interventions (including VRET, VR Cognitive Behavioral Therapy, VR Occupational Therapy, VR Skill Training) and found them to be effective when compared to both waitlist and non-VR controls (though more effective when compared to the former), and considers that VR has potential to be used in more than only exposure therapy.

Moderator analysis was performed in most of the meta-analyses (DENG et al., 2019; CARL et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; TURNER; CASEY, 2014). The most commonly explored moderators were number of sessions (DENG et al., 2019; CARL et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017), sample size (CARL et al., 2019; CARDOŞ; DAVID; DAVID, 2017), publication year (CARL et al., 2019), the disorder being treated (BENBOW; ANDERSON, 2019), demographics (DENG et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; TURNER; CASEY, 2014), combination to other treatment (BENBOW; ANDERSON, 2019), the use of “homework” (BENBOW;

ANDERSON, 2019), study quality (CARDOŞ; DAVID; DAVID, 2017), outcome type (CARDOŞ; DAVID; DAVID, 2017), length of follow-up (CARDOŞ; DAVID; DAVID, 2017), and type of VR intervention and control (TURNER; CASEY, 2014). Among these, number of sessions (DENG et al., 2019) and the use of "homework" (BENBOW; ANDERSON, 2019), had positive effects; while participants' age (CARDOŞ; DAVID; DAVID, 2017), and study quality (CARDOŞ; DAVID; DAVID, 2017) had inverse effects (i.e., younger participants had more pronounced effects from the intervention, lower quality studies showed bigger effects). Cardoş, David and David (2017) explain the moderation effect of age by arguing that younger participants, being consumers of new technology, can adapt more easily to VR.

6.3.4 What are the criticisms to the primary studies?

The main criticisms to the primary studies are methodological, such as high or unclear risk of bias (GUJJAR et al., 2019; FERNÁNDEZ-ÁLVAREZ et al., 2019; DENG et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; TURNER; CASEY, 2014), poor availability of primary data and methodological details (FERNÁNDEZ-ÁLVAREZ et al., 2019; DENG et al., 2019; BENBOW; ANDERSON, 2019; KOSKINA; CAMPBELL; SCHMIDT, 2013), small sample sizes (FERNÁNDEZ-ÁLVAREZ et al., 2019; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; TURNER; CASEY, 2014), too specific/homogeneous samples (DENG et al., 2019; CARDOŞ; DAVID; DAVID, 2017), and no study on long-term effects (GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018; CARDOŞ; DAVID; DAVID, 2017). In terms of intervention design, the main criticisms were insufficient exposure time (CARDOŞ; DAVID; DAVID, 2017; KOSKINA; CAMPBELL; SCHMIDT, 2013) and the mixing of intervention protocols (KOSKINA; CAMPBELL; SCHMIDT, 2013).

6.3.5 What are the paths for future research for the authors in this area?

Authors argue that a different set of outcomes is necessary: safety, acceptability, attendance (GUJJAR et al., 2019); and cost-effectiveness (GUJJAR et al., 2019; MARTIN et al., 2011); more research on adverse effects (FERNÁNDEZ-ÁLVAREZ et al.,

2019; DIAS; BARBOSA; VIANNA, 2018); effectiveness should be analyzed on several levels (cognitive, emotional, behavioral, psychophysiological) (CARDOŞ; DAVID; DAVID, 2017) as well as in secondary outcomes (depression and quality of life), and use other methods to complement self-reports (KAMPMANN; EMMELKAMP; MORINA, 2016), such as eye tracking, startle response and functional magnetic resonance imaging (KOSKINA; CAMPBELL; SCHMIDT, 2013). In three surveys (GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018; CARDOŞ; DAVID; DAVID, 2017; TURNER; CASEY, 2014) it was noted that immersion level could be a possible moderator, but also that it was poorly reported as a variable in the primary studies. Turner and Casey (2014) suggest that presence should be measured in future studies, and that the studies should report more details of the VEs used (such as providing a screenshot), and that information on the digital literacy of the participants should also be reported. Benbow and Anderson (2019) cite failure to immerse in the virtual environment as the most prominent VRET-specific reason for drop-out, and they argue that finding ways to increase immersion and limit side-effects should be a path for future research – which also involves reporting such variables, something that most of the primary studies did not do. Ghiţă and Gutiérrez-Maldonado (2018) suggest that future research should systematically vary the level of immersion and investigate its effects and side-effects – they argue that the increasing immersion might not necessarily increase effectiveness, but might bring greater side-effects such as cybersickness. Finally, Koskina, Campbell and Schmidt (2013) also suggest future research should incorporate advances in learning theory and the underlying mechanisms of extinction and reconsolidation (in the case of eating disorders).

6.4 Post-stroke Rehabilitation

Eleven studies focused on physical rehabilitation after stroke (CALDERÕN et al., 2019; PERROCHON et al., 2019; MOHAMMADI et al., 2019; LIN et al., 2019; CHEN et al., 2019; SUBRAMANIAN; PRASANNA, 2018; RAVI; KUMAR; SINGHI, 2017; HOWARD, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016; KLINKE et al., 2015; CORBETTA; IMERI; GATTI, 2015). Most of them evaluated XR in terms of its effectiveness in reducing patients' symptoms. One survey addressed rehabilitation of cerebral palsy rather than stroke (RAVI; KUMAR; SINGHI, 2017).

6.4.1 What is XR for this category of studies?

All of the 11 secondary studies for physical rehabilitation focused on VR. Three studies did not define VR (CALDERÕN et al., 2019; SUBRAMANIAN; PRASANNA, 2018; KLINKE et al., 2015). The definition of VR in the remaining studies was of a computer-generated environment that allows for interaction.In practice, the included primary studies were mainly screen-based and used commercial exercise games (such as the Microsoft Kinect and Nintendo Wii, or the more specific Interactive Rehabilitation Exercise Software – IREX), while only one study reported the inclusion of HMDs (CORBETTA; IMERI; GATTI, 2015). According to (PERROCHON et al., 2019), this lack can be attributed to the high cost of the HMDs.

6.4.2 What's the motivation for the use of VR?

The main justifications for the use of VR in rehabilitation were increased patient motivation, real-time feedback, the encouragement of a high number of repetitions (high intensity), VR's suitability for task-oriented training and its adaptability.

Motivation is considered a promoter of motor learning (SUBRAMANIAN; PRASANNA, 2018; RAVI; KUMAR; SINGHI, 2017; HOWARD, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016; CORBETTA; IMERI; GATTI, 2015), and VR is said to increase motivation through the use of gaming and competition elements (PERROCHON et al., 2019; CHEN et al., 2019; HOWARD, 2017), or simply novelty (HOWARD, 2017).

Real-time feedback is another promoter of motor learning (CALDERÕN et al., 2019; SUBRAMANIAN; PRASANNA, 2018; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016; CORBETTA; IMERI; GATTI, 2015) afforded by VR via observing the avatar performance on-screen or in the form of game points.

Intensity (generally in terms of number of exercise repetitions) is also a promoter of motor learning (CALDERÕN et al., 2019; PERROCHON et al., 2019; SUBRAMANIAN; PRASANNA, 2018; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016). VR is said to encourage the patient to perform a high number of exercise repetitions (intensity), without exhausting or boring the user, which can be linked to motivation. Other ways VR can help increase repetitions is through variation (e.g., different games) (LIN et al., 2019; CHEN et al., 2019; SUBRAMANIAN; PRASANNA, 2018).

Task-oriented training is a rehabilitation approach that employs exercises that

mimic functional movements of daily living (e.g., walking, picking up a phone). VR is found suitable for task-oriented training (CALDERÕN et al., 2019; PERROCHON et al., 2019; CORBETTA; IMERI; GATTI, 2015) due to its ability to simulate real environments. On the other hand, results from a survey (HOWARD, 2017) show that this physical and cognitive fidelity afforded by VR is not yet a proven mediator of rehabilitation, thus deeming further research necessary (i.e., studying varying levels of physical fidelity and their impact on rehabilitation outcomes).

VR's adaptability to specific user needs is mentioned as an advantage (LIN et al., 2019). However, in two surveys (PERROCHON et al., 2019; RAVI; KUMAR; SINGHI, 2017) authors argue that dropouts from the interventions were often caused by the unsuitability of (off-the-shelf) VR systems for the purposes of rehabilitation. Perrochon et al. (2019) suggest that researchers should study the acceptability and feasibility of the interventions, and Ravi, Kumar and Singhi (2017) argue that dropouts could be diminished by designing the interventions taking into account users' opinions. They also argue that the cognitive capabilities required by the interventions limit its applicability in some cases for participants with cerebral palsy.

6.4.3 What did the evaluations on the primary studies focus on?

All of the studies were concerned with effectiveness of VR interventions in patient outcomes (such as balance and motor control). This was mostly measured through clinical tests and scales specific to the domain area, such as the Fugl-Meyer Assessment and the Berg Balance Scale, which are observation-based measures. Other reported measures of these outcomes included the use of instrumentation such as computerized posturography, force plates and surface EMG. One of the studies (CHEN et al., 2019) also included outcomes on the feasibility of the usage of VR interventions at home (in terms of technical barriers and user motivation).

Although increased motivation is cited as one of the reasons for the application of VR in rehabilitation in eight studies (PERROCHON et al., 2019; LIN et al., 2019; CHEN et al., 2019; SUBRAMANIAN; PRASANNA, 2018; HOWARD, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016; CORBETTA; IMERI; GATTI, 2015) it only appeared as outcomes in two studies (CHEN et al., 2019; RAVI; KUMAR; SINGHI, 2017). Chen et al. (2019) used a self-report inventory, while Ravi, Kumar and Singhi (2017) did not provide a measure. Corbetta, Imeri and Gatti (2015) see this gap in measures of motiva-

tion as an issue of the primary studies. They point out that the subjective preferences and attitude of users were not considered in the studies, but are relevant for compliance to the intervention.

Six of the secondary studies were meta-analyses (PERROCHON et al., 2019; MOHAMMADI et al., 2019; LIN et al., 2019; SUBRAMANIAN; PRASANNA, 2018; HOWARD, 2017; CORBETTA; IMERI; GATTI, 2015). In the first (PERROCHON et al., 2019), exercise-based games (EBGs) were not found superior to traditional interventions, while having a higher drop-out rate. The main reason for dropout was the belief that EBGs could increase risk factors. The other three surveys reported reasons for drop-out related to the off-the-shelf games characteristics (lack of customization, childish game design, lack of accessibility to the technology). Mohammadi et al. (2019) analyzed balance outcomes and found that VR in addition to conventional therapy increased balance moderately compared to only conventional therapy. Other authors (LIN et al., 2019) also did not find significant superiority of VR over conventional therapy in lower extremity motor recovery, and explain this finding as due to VR and conventional rehabilitation being based on the same mechanisms of motor learning and use-dependent theory – thus it would be the task (similar in both interventions), and not the modality, that was responsible for the outcomes. Subramanian and Prasanna (2018) analyzed the use of VR as a complementary therapy to non-invasive brain stimulation, and found preliminary evidence encouraging its use for motor improvement post-stroke. In another study (HOWARD, 2017) virtual reality rehabilitation (VRR) was found to have moderate positive effects compared to active controls, when analyzing motor control, balance, gait and strength together (though when analyzed individually, some of these outcomes were not significant). The authors conclude that the benefits of VRR are proven, but its cost might not be justified, thus requiring a cost-benefit analysis. Corbetta, Imeri and Gatti (2015) found that VR based rehabilitation for walking speed, balance and mobility was found more effective than standard rehabilitation of the same duration, but when used as an addition to standard rehabilitation (increasing total duration), no significant effect was found. In both cases, the effects were smaller than the “smallest real difference” of the scales used, thus their clinical relevance is questioned.

6.4.4 What are the criticisms to the primary studies?

Several methodological issues were pointed out by the authors: small samples (MOHAMMADI et al., 2019; RAVI; KUMAR; SINGHI, 2017; HOWARD, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016; KLINKE et al., 2015), lack of standardized outcomes (HOWARD, 2017; KLINKE et al., 2015), no follow-up (RAVI; KUMAR; SINGHI, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016), variability of patient characteristics (RAVI; KUMAR; SINGHI, 2017; KLINKE et al., 2015), non-comparable intervention and control (HOWARD, 2017) and possible bias (MOHAMMADI et al., 2019). Heterogeneity across the different treatment protocols was pointed out in three surveys(CALDERÕN et al., 2019; HOWARD, 2017; KLINKE et al., 2015). Mohammadi et al. (2019) criticise the primary studies for not describing the games (or non-game content) used and their characteristics (e.g., competitive or team-based), as well as the other components of the VR system and intervention overall.

Poor choice of outcomes was also found as an issue (SUBRAMANIAN; PRASANNA, 2018; RAVI; KUMAR; SINGHI, 2017; CORBETTA; IMERI; GATTI, 2015). For example, Corbetta, Imeri and Gatti (2015) advise for the investigation of participants' attitude; Subramanian and Prasanna (2018) criticize the use of time as a metric of intensity rather than repetitions, and argue that the clinical outcomes do not distinguish between behavioral recovery and compensation (i.e., muscle substitutions). They further suggest the use of kinematic measures as a more appropriate outcome. Ravi, Kumar and Singhi (2017) argue that future research should develop clinically validated scales in the virtual environment rather than relying on standard game scores, for daily quantification of improvement. Other additional outcomes suggested are ease of use, acceptability and feasibility by the patient and relatives or caregivers and monitoring compliance (PERROCHON et al., 2019), and adverse effects (MOHAMMADI et al., 2019; RAVI; KUMAR; SINGHI, 2017). The addition of retention tests of long-term effects was also suggested (MOHAMMADI et al., 2019).

6.4.5 What are the paths for future research for the authors in this area?

Several paths of future research are open in this area, such as investigating the possible role of immersion and presence(RAVI; KUMAR; SINGHI, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016), and finding the optimal dosage and characteristics of the

VR intervention (MOHAMMADI et al., 2019; CORBETTA; IMERI; GATTI, 2015) by comparing different interfaces (VIÑAS-DIZ; SOBRIDO-PRIETO, 2016) and levels of physical fidelity (HOWARD, 2017), for example.

Changes to intervention design are also advised, such as designing interventions for engagement (CHEN et al., 2019) and multi-player use (PERROCHON et al., 2019), while accounting for its acceptability and suitability to the patient's setting (PERROCHON et al., 2019; CHEN et al., 2019). Other than accounting for the home environment and the user's technical abilities (CHEN et al., 2019), research on how the participant experiences the intervention is also warranted (KLINKE et al., 2015) (this can also be linked to the lack of research on patients' subjective preference and attitude (CORBETTA; IMERI; GATTI, 2015)). Finally, the design of subject-specific VR systems was also suggested (RAVI; KUMAR; SINGHI, 2017), in contrast to relying solely on off-the-shelf game systems.

Investigating the cost-effectiveness of VR rehabilitation is a suggested path (PERROCHON et al., 2019; HOWARD, 2017). Others (HOWARD, 2017; CORBETTA; IMERI; GATTI, 2015) suggest that the sources of heterogeneity in their meta-analyses should be further investigated.

A seemingly dissonant conclusion is related to suggestions for future study designs: one of the surveys (PERROCHON et al., 2019) recommend the addition of passive control groups to distinguish between the effects of VR and the effects of simply practicing the tasks, while other (LIN et al., 2019) argue that new studies should focus on superiority against conventional therapies, and not solely on effectiveness against no intervention. Both views can meet on an experiment design that includes both active and passive controls, or, in case of a well-established "gold standard" therapy of proven effectiveness, an experiment design that aims to prove superiority of VR over it.

Finally, in terms of the mechanisms of VR rehabilitation, Howard (2017) suggest that the mediation role of motivation on outcomes should be investigated, while other authors (VIÑAS-DIZ; SOBRIDO-PRIETO, 2016) suggest that future research should focus on understanding if VR affects cortical reorganisation.

6.5 Cognition

Eight studies dealt with cognition and the use of VR as a treatment (MORENO et al., 2019; ALASHRAM et al., 2019; TAN; LEE; LEE, 2018; COYLE; TRAYNOR;

SOLOWIJ, 2015) or as a diagnosis tool (PEREIRA et al., 2019; BLUETT; BAYRAM; LITVAN, 2019; COGNÉ et al., 2017) for neurocognitive disorders. Lastly, one study (NEGUŢ et al., 2016) investigated VR in cognitive assessments in general.

6.5.1 What is XR for this category of studies?

All but two studies (BLUETT; BAYRAM; LITVAN, 2019; TAN; LEE; LEE, 2018) explicitly defined VR. VR is viewed as a computer-generated 3D environment (PEREIRA et al., 2019; MORENO et al., 2019; ALASHRAM et al., 2019; TAN; LEE; LEE, 2018; COGNÉ et al., 2017; NEGUŢ et al., 2016) that can be interacted with (PEREIRA et al., 2019; MORENO et al., 2019; ALASHRAM et al., 2019; TAN; LEE; LEE, 2018; NEGUŢ et al., 2016). For two surveys (COGNÉ et al., 2017; NEGUŢ et al., 2016), this artificial environment is close to reality. Immersion in VR gives the user a sense of presence in the virtual environment (MORENO et al., 2019; NEGUŢ et al., 2016), and different levels of immersion can be provided (e.g., fully immersive and non-immersive (MORENO et al., 2019)), which depends on the display and interaction devices used (e.g., HMDs and gloves). As for AR, it is a mixture of real and virtual worlds (along the reality-virtuality continuum), that can have virtual objects overlaid on top of the real world (PEREIRA et al., 2019). Moreno et al. (2019) classified the applications of the primary studies in *subjective* and *objective* levels of immersion, the latter using five factors (inclusiveness, extensiveness, surrounding, vividness and matching) as proposed by Slater and Wilbur (1997). However, nearly one fifth of the included primary studies did not provide enough information to carry out the objective classification.

6.5.2 What's the motivation for the use of XR?

The main motivation for the usage of VR was its ecological validity (PEREIRA et al., 2019; BLUETT; BAYRAM; LITVAN, 2019; COGNÉ et al., 2017; NEGUŢ et al., 2016; COYLE; TRAYNOR; SOLOWIJ, 2015). VR could be used to trigger Parkinson's disease freezing of gait symptoms in a safe and controlled manner, with the subject seated down (PEREIRA et al., 2019; BLUETT; BAYRAM; LITVAN, 2019). A further advantage of this setting is that brain imaging can be recorded, thus potentially allowing to elucidate the mechanism of the condition in a way that's not normally feasible in the real

environment. Cogné et al. (2017) also cited the safety of VR and that learning and finding trajectories relies on the same mechanisms for both real and virtual environments, thus allowing researchers to test disabilities more easily than on real environments. For other authors (NEGUŢ et al., 2016; COYLE; TRAYNOR; SOLOWIJ, 2015), this ecological validity meant that training tasks and assessments performed in VR could be better generalized to the real scenarios. The use of HMDs to achieve a higher level of immersion is explicit in two surveys (NEGUŢ et al., 2016; COYLE; TRAYNOR; SOLOWIJ, 2015). However, in the scope of treatment of neurocognitive disorders, Moreno et al. (2019) argue that there is no evidence that higher immersion leads to better effects.

Other reasons for the use of VR are cost-effectiveness (COGNÉ et al., 2017; COYLE; TRAYNOR; SOLOWIJ, 2015), motivation and the potential of unlimited repetition of training tasks (ALASHRAM et al., 2019)

6.5.3 What did the evaluations on the primary studies focus on?

Four of the studies reported outcomes of cognitive performance, measured with domain-specific scales (such as the Fuld Object-Memory Evaluation) (MORENO et al., 2019; ALASHRAM et al., 2019; TAN; LEE; LEE, 2018; COYLE; TRAYNOR; SOLOWIJ, 2015). These assessments sought to find the effects of VR in the treatment or rehabilitation of patients with neurocognitive disorders.

In a meta-analysis (NEGUŢ et al., 2016), the cognitive assessment itself was being studied. More specifically the impact of the virtual environment on participants' scores: the meta-analysis compared the performance on cognitive assessments when they were administered via pen and paper, computers, or VR. Age, gender, clinical status, type of control and type of task (time-based or error-based) were treated as potential moderators. VR assessments led to poorer performance, especially in executive function measures, and mixed results in memory measures. Age was a significant moderator in decreasing the effect size (i.e., older participants had a more similar performance across types of assessment); clinical status was also significant, with healthy participants having larger effect sizes; finally, type of task also moderated the effect, with time-based measures showing a larger penalty in VR than error-based ones. Overall, the authors conclude that cognitive performance in VR is poorer than on paper-based or computerized assessments, which might be due to the increased level of complexity and difficulty of VR. The lower scores achieved using VR were seen as an advantage, since they pointed towards poten-

tially higher ecological validity.

The other three studies (PEREIRA et al., 2019; BLUETT; BAYRAM; LITVAN, 2019; COGNÉ et al., 2017) used the virtual environment as a platform for the diagnosis of disorders. The participants had their step latency measured, as an indicator of the presence of Parkinson's disease freezing of gait while walking around the virtual environment (BLUETT; BAYRAM; LITVAN, 2019). This measure was triangulated with self-reports of symptoms as well as functional magnetic resonance imaging (fMRI). VEs were used to diagnose spatial navigation disorders (COGNÉ et al., 2017), through the performance of users in virtual navigation tasks. VR was also used to study different designs of navigational aids, both to understand how people navigate as well as to help design better (real) environments.

6.5.4 What are the criticisms to the primary studies?

The main methodological criticisms are the lack of a clear clinical description of the participants (MORENO et al., 2019; BLUETT; BAYRAM; LITVAN, 2019), lack of randomized controlled trials (MORENO et al., 2019; ALASHRAM et al., 2019), and small sample sizes (ALASHRAM et al., 2019; COYLE; TRAYNOR; SOLOWIJ, 2015).

In terms of outcomes, Bluett, Bayram and Litvan (2019) criticize the use of arbitrary measures in a non-validated test, while Coyle, Traynor and Solowij (2015) criticize the reliance on self-reports.

6.5.5 What are the paths for future research for the authors in this area?

Moreno et al. (2019) argue that user acceptance and adverse effects should be assessed systematically. They also ask for the comparison of different levels of immersion on the effectiveness of treatment. Further studies should be longitudinal and investigate the generalization of the benefits of VR. Step latency is an arbitrary measure for freezing of gait, and future research should aim to validate the VR walking course by comparing it to a real one (BLUETT; BAYRAM; LITVAN, 2019). (COGNÉ et al., 2017) asks for more research on the impact of the characteristics of navigational cues and the underlying cognitive processes.

For Neguţ et al. (2016), future studies should focus on the predictive validity of

the cognitive assessments. (COYLE; TRAYNOR; SOLOWIJ, 2015) asks for the inclusion of both active and waitlist controls in future studies, to differentiate condition and placebo effects. Also, investigate whether VR cognitive training is more effective for a particular diagnostic group, employing functional outcomes (such as observation and performance based), rather than only self-reports. Future research should help develop an understanding of the underlying neurobiology involved with cognitive training, and be longitudinal.

6.6 Surgery

Six studies focused on the use of XR as an aid for surgery (DETMER et al., 2017; JODA et al., 2019; BOSC et al., 2019; BERNHARDT et al., 2017; GRECO et al., 2014; KERSTEN-OERTEL; JANNIN; COLLINS, 2013). Evaluations mostly concerned the technical performance of the system.

6.6.1 What is XR for this category of studies?

Two studies (DETMER et al., 2017; JODA et al., 2019) included both VR and AR studies. VR is defined as a computer generated simulation that allows the immersion and interaction with an artificial 3D environment (although Detmer et al. (2017) state that the immersive character of VR, e.g., using HMDs, was not considered mandatory for the inclusion of primary studies).

In three surveys (DETMER et al., 2017; JODA et al., 2019; BOSC et al., 2019), AR is defined as a technology that superimposes virtual objects into the real world. While an interaction aspect of AR is cited as fundamental (JODA et al., 2019), two studies (DETMER et al., 2017; BOSC et al., 2019) focus more on the visualization aspect. Other was concerned with the time for overlaying virtual objects (BOSC et al., 2019). Two reviews (GRECO et al., 2014) included AR studies, but did not provide a definition. Finally, two last surveys (BERNHARDT et al., 2017; KERSTEN-OERTEL; JANNIN; COLLINS, 2013) use the continuum definition of Mixed Reality (MILGRAM; KISHINO, 1994), and thus consider augmented reality and augmented virtuality as points along this continuum.

Other applications (BOSC et al., 2019; BERNHARDT et al., 2017; KERSTEN-

OERTEL; JANNIN; COLLINS, 2013) were described in terms of diverse characteristics of the systems: tracking and registration methods, type of display (HMD, projection on the patient, half-silvered mirror, handheld). The remaining studies did not provide a description of the included systems.

6.6.2 What's the motivation for the use of XR?

The main reason for the use of AR and MR cited in the studies is the ability to incorporate additional information in the operative field of view, cited by all studies. In minimally invasive (laparoscopic) surgery, AR and MR can allow the visualization of anatomical structures in place (e.g., superimposed onto the patient's body), without having to divert their attention to an external monitor for the endoscope's video feed (BERNHARDT et al., 2017; GRECO et al., 2014; KERSTEN-OERTEL; JANNIN; COLLINS, 2013). Another advantage afforded by this superimposition is the enhancement of the surgeon's spatial orientation compared to the 2D endoscope view alone (DETMER et al., 2017; BERNHARDT et al., 2017). It also allows for visually synthesizing diverse sources of preoperative information (BOSC et al., 2019; BERNHARDT et al., 2017), which has the potential for decreasing the surgeon's cognitive load (BERNHARDT et al., 2017). On the other hand, it was also argued that AR overlays can actually induce inattentional blindness (DETMER et al., 2017).

The use of VR is mainly motivated by the capacity of simulating procedures (DETMER et al., 2017; JODA et al., 2019). This affords unlimited practice time (JODA et al., 2019) and precise planning of surgical interventions (DETMER et al., 2017).

6.6.3 What did the evaluations on the primary studies focus on?

Accuracy of the overlaid images was the primary outcome of the applications in all six studies. This was evaluated by measuring the error of the superimposition in relation to a known anatomical landmark, in a phantom or on a real patient (BOSC et al., 2019). However, the measurement of accuracy is still an issue, since it is hard to check if the augmentation is placed correctly when it is deep below the skin level (BERNHARDT et al., 2017). Some studies that evaluated the perceived usability and usefulness of the applications via qualitative feedback from the users were also included in two surveys

(DETMER et al., 2017; KERSTEN-OERTEL; JANNIN; COLLINS, 2013).

6.6.4 What are the criticisms to the primary studies?

The main criticism of the primary studies was the focus on technical validation (i.e., accuracy), or qualitative evaluation (perceived usefulness) as the main outcomes, rather than actual patient outcomes (such as blood loss, complications, length of hospital stay) that are also indicators of the effectiveness of the XR systems (DETMER et al., 2017; BERNHARDT et al., 2017; KERSTEN-OERTEL; JANNIN; COLLINS, 2013).

For Bosc et al. (2019), there is a disparity in the description of the AR systems and their use, making them hard to compare. Also, comparative studies with reference procedures are lacking.

6.6.5 What are the paths for future research for the authors in this area?

Larger clinical studies are suggested (DETMER et al., 2017; JODA et al., 2019; BOSC et al., 2019) as a way to demonstrate the effectiveness and cost-effectiveness of XR applications. Future research should consider how intuitive and easy to use the systems are (DETMER et al., 2017). These authors suggest the use of familiar interfaces for the surgeon (such as a tablet), rather than unfamiliar ones (such as a 3D mouse). They also outline the need to take human factors and HCI methods into account when designing the system. The systems should also allow the visualization of uncertainty (of segmentation, registration and tracking) during the operation. New models that account for real-time deformation and movement of organs are also necessary. Some surveys (BERNHARDT et al., 2017; KERSTEN-OERTEL; JANNIN; COLLINS, 2013) report that little research focused on how the surgeon perceives the AR systems, both literally in terms of visual perception (i.e., depth perception) and also in terms of what information is most useful at any given moment. Usability is often deemed secondary and not evaluated in the primary studies, but accuracy alone is not enough to ensure the utility of AR in terms of clinical outcomes, and further multidisciplinary research (i.e., medicine and engineering) is needed (BERNHARDT et al., 2017). Human factors and psychophysical measures of visualization and interaction methods in MR surgery are lacking (KERSTEN-OERTEL; JANNIN; COLLINS, 2013). They suggest researchers include user studies to learn the

needs and constraints of surgeons in the operating room to improve the design of XR applications.

6.7 Pain relief

Five studies focused on the treatment of pain using virtual reality (CHI et al., 2019; SCHEFFLER et al., 2018; SCAPIN et al., 2018; CÁRDENAS; ARANDA, 2017; MALLOY; MILLING, 2010). The main goal of the studies was to investigate the effectiveness of VR in relieving diverse types of pain. In two studies (SCHEFFLER et al., 2018; CÁRDENAS; ARANDA, 2017), VR interventions are studied alongside other types of interventions.

6.7.1 What is XR for this category of studies?

All four studies that defined VR (CHI et al., 2019; SCHEFFLER et al., 2018; SCAPIN et al., 2018; MALLOY; MILLING, 2010) consider it a 3D virtual environment that allows for immersion. The use of multiple senses is also suggested (SCAPIN et al., 2018; MALLOY; MILLING, 2010) as a definition. All four studies included HMDs. In Chi et al. (2019) "immersive VR" means that HMDs are used while other display methods that do not block the perception of the real world as much are "non-immersive VR". Two surveys (SCHEFFLER et al., 2018; MALLOY; MILLING, 2010) included HMD studies exclusively and seems to equate them to VR, but the latter includes studies comparing HMDs and computer screens. Scapin et al. (2018) includes both HMDs and off-the-shelf video-games, while the fifth survey (CÁRDENAS; ARANDA, 2017) presents two studies on VR Mirror Visual Feedback. Interaction with the virtual world was present on all five studies; in one of them (MALLOY; MILLING, 2010) the presence of interaction in at least one of the VR interventions of each primary study was an inclusion criteria.

6.7.2 What's the motivation for the use of XR?

The main reasons cited for employing VR in pain relief are: the ability of VR to act as a distraction and thus reduce the perception of pain, and the minimal side-effects of VR (compared to pharmacological interventions).

All five studies agree that VR can be used to distract patients suffering from pain. They argue that distraction reduces pain by seizing cognitive resources that would otherwise be used to process the pain signals. The distracting effects might be due to the afforded immersion itself (SCAPIN et al., 2018; MALLOY; MILLING, 2010) or because of game mechanics that elicit active cognition from the participant (MALLOY; MILLING, 2010). Thus, the authors argue that both the role of immersion and of fun of the virtual environment should be further studied as both can moderate distraction and thus pain reduction. Measures of perceived fun are present in one survey (SCAPIN et al., 2018), while in another (CÁRDENAS; ARANDA, 2017), interaction itself is viewed as a motivating factor for the use of VR, since it makes the subjects feel committed and rewarded, motivating them to repeat the exercises with greater intensity (in this case, the participants were undergoing therapy for phantom limb pain using VR-based mirror visual feedback, a part of which involves exercises with the surviving limb). Scapin et al. (2018) also considers VR as a promoter of treatment adherence. For Chi et al. (2019), in addition to capturing attention, VR can modulate the perception of pain by influencing the user's emotion.

The non-pharmacological, non-invasive nature of VR is viewed as an advantage over the usual pharmacological treatments for pain (CHI et al., 2019; SCHEFFLER et al., 2018). As for the possible side-effects of VR, such as cybersickness, Scapin et al. (2018) note that nausea is a common side-effect of analgesics for burn patients – what confounds the cause of nausea, while Chi et al. (2019) noted distress, musculoskeletal pain and fatigue in some cases, but that could be caused by increased activity overall and not specifically by VR.

Portability, customization and ease of use are also cited (CHI et al., 2019). On the other hand, the size of VR devices and the lack of knowledge of health professionals about them as barriers for their adoption are also cited as motivation (SCHEFFLER et al., 2018). Interestingly, cost-effectiveness is cited as an advantage of VR by the earliest study in the group (MALLOY; MILLING, 2010) – from 2010 – which considered the cost of VR headsets at the time (2,000 to 3,000 US dollars) promising for wider adoption.

6.7.3 What did the evaluations on the primary studies focus on?

All five studies included outcomes of pain, measured using pain scales and questionnaires, such as the Visual Analog Scale (CHI et al., 2019; SCAPIN et al., 2018), or re-

ported time spent thinking about pain (SCHEFFLER et al., 2018; MALLOY; MILLING, 2010). Other methods for measuring pain included physiological measures (heart rate (SCAPIN et al., 2018; MALLOY; MILLING, 2010), salivary cortisol and magnetic resonance imaging (SCAPIN et al., 2018)) and observation of patient behavior (e.g., the frequency of scratching (MALLOY; MILLING, 2010)). Secondary outcomes such as anxiety were also included (SCHEFFLER et al., 2018; SCAPIN et al., 2018). Scapin et al. (2018) aggregated studies with the most variation in measured outcomes, such as fun, side-effects, perceived game quality, human resources needed for the intervention and cost, as well as measures of assembly and cleaning time of VR equipment. Scheffler et al. (2018) is a meta-analysis of several non-pharmacological interventions for relief of pain in burn patients. VR was found to have the largest effect compared to passive and attention controls. The authors found significant heterogeneity in the VR trials. VR games and VR content delivery (i.e., relaxation DVD watched through an HMD) were bundled together as VR.

6.7.4 What are the criticisms to the primary studies?

Methodological issues are a main criticism (including risk of bias and overall study quality (CHI et al., 2019; SCHEFFLER et al., 2018; SCAPIN et al., 2018; CÁRDENAS; ARANDA, 2017) and small sample sizes (SCHEFFLER et al., 2018; CÁRDENAS; ARANDA, 2017; MALLOY; MILLING, 2010)).

The inclusion of only pain outcomes is criticized (CHI et al., 2019; MALLOY; MILLING, 2010). The former (CHI et al., 2019) suggests the usage of a standardized set of outcomes that also considers physical and emotional functioning as well as patient ratings of satisfaction and improvement; the latter (MALLOY; MILLING, 2010) asks for presence and fun of the virtual environments to be evaluated, so that they can be explored as possible moderator variables; it also questions the usage of unreliable behavioral metrics (such as a pain rating by the patient's parents).

6.7.5 What are the paths for future research for the authors in this area?

The question of whether immersion moderates the effect of VR in pain reduction (e.g., by comparing immersive and non-immersive VR systems) is cited as a direction

for future research in both the earliest and the latest studies in the group (MALLOY; MILLING, 2010; CHI et al., 2019). The analgesic effect provided by VR is still not fully understood (CHI et al., 2019; SCHEFFLER et al., 2018; MALLOY; MILLING, 2010). The cost-effectiveness of VR should be studied by comparing it to other types of treatment (MALLOY; MILLING, 2010) and reporting on all the costs and personnel involved in the delivery of the VR intervention (SCAPIN et al., 2018).

Finally, Scapin et al. (2018) asks for a more standardized definition of VR interventions, and Malloy and Milling (2010) suggests that studies should be mindful of the virtual environments used, and design for increased presence and fun (for example, the authors argue that exploring a virtual kitchen could be considered less fun than playing a game where you throw snowballs at virtual foes).

The next three categories contain five, three and two papers, respectively. Therefore, we change our narrative synthesis to briefly summarise each paper.

6.8 Physical prevention

Five surveys focused on the use of XR as a means to increase physical activity in healthy adults (VOGT et al., 2019; NG et al., 2019; CACCIATA et al., 2019; MARTINS et al., 2018; NEUMANN et al., 2018). We summarise each of them below. The first Vogt et al. (2019) is a survey of studies on VR for balance prevention and rehabilitation. It included 11 studies on balance prevention on healthy adults and 5 studies on balance rehabilitation after lower limb impairment. The main driver for the use of VR is motivation, which is believed to lead to higher compliance to the intervention. This motivation can be garnered through the use of exciting virtual environments for performing tasks, and the use of gaming and competition elements. Other potential advantages of VR are its ability to provide real-time, continuous feedback, since that can be a promoter of motor learning; and VR's adaptability to specific user needs. Most of the primary studies used off-the-shelf video game systems, such as the Nintendo Wii and Xbox Kinect. None made use of HMDs. The authors suspect that might be due to three reasons: (i) the relative novelty of the technology and lack of available balance games; (ii) HMDs can alter head position simply by being worn, thus influencing balance and negatively affecting its usefulness for balance training; and (iii) cybersickness due to the use of HMDs.

The authors criticize the choice of outcomes in the primary studies, since several different methods were used to measure static and dynamic balance, making it hard to

compare them directly. The same is true for the variability in the implementation of the interventions. Finally, only 4 studies included a control group that effectively received no balance treatment, making them more likely to be able to discriminate between the real effect of an intervention and the effect of practicing the balancing task itself. Despite the argument for motivation as a potential advantage of VR, the authors note that the subjective perspective of the participants about the VR intervention was not considered in the review. However, they argue it is an important factor to determine the compliance of a therapy, and thus should be further investigated.

Recommendations for future research are: (i) adopting a clearer, systematized definition of VR; (ii) investigate the use of newer technologies such as HMDs, as well as investigating its effects not only on a behavioral level, but also in a neurophysiological level, when compared to traditional balance interventions; (iii) selecting a "gold-standard" measure for balance outcomes, to make comparison between studies easier; and (iv) strengthening the study designs by making sure the VR and non-VR conditions are comparable, including a passive control group and performing retention tests long after the post-test.

A meta-analysis (NG et al., 2019) sought to find evidence of effectiveness of VR and AR interventions on physical activity (broken down into frequency, intensity and duration of exercise), physical performance and psychological outcomes in healthy subjects. It included 22 studies. The main reasons stated for the use of XR on exercise are its immunity to weather, light and traffic conditions compared to traditional outdoors exercising, the motivation elicited by XR, and the inability of the human brain to distinguish real and virtual stimuli.

VR and AR are placed on the reality-virtuality continuum. VR is defined as an interactive digital environment that tracks users activities. AR is defined as a mixture of real and virtual environments. The systems presented in the primary studies varied from exercise video games (Microsoft Kinect and Nintendo Wii Fit) to VR biking, dancing and treadmill walking. The applications were classified as immersive when a HMD was used and as non-immersive when a 2D screen was used as display.

The outcomes of physical activity, physical performance and psychological outcomes were measured mostly through observation-based tests (such as the Timed Up and Go test), as well as the attendance of exercise sessions and instrumented measures such as surface EMG.

In the meta analysis, possible moderators were immersiveness (i.e., employed an HMD or not), type of reality (VR or AR), year of publication, percentage of females,

mean age of the sample, intervention duration, number of sessions, and minutes per session. The results were a large effect found on physical activity, a small one on physical performance, and no effect on psychological outcomes. No moderator was found significant. The authors criticize the lack of a theory-testing approach on the primary studies (of e.g., health-belief models or motor learning theory). They suggest that the effects can be explained by two mechanisms of social cognitive theory: (i) vicarious reinforcement (i.e., interventions where participants observed and followed avatars in the VE); and (ii) identification (or the lack thereof, since the avatars were not personalised for the participants); thus, they suggest future studies should use the trans-theoretical model and investigate the effects of virtual representation of self. They also argue future studies should include measures of presence and of previous exercising behavior to be used as possible moderators. They should also perform RCTs (especially for immersive VR and AR, which were the least common interventions) and employ multi-armed factorial designs to devise which characteristics of XR are responsible for intervention efficacy.

Cacciata et al. (2019) analyzed nine RCTs of exergaming interventions for older adults. They tried to find a link between exergaming and increases in quality of life. VR games are considered one type of exergame, that monitors body movement and provides real time feedback. Most of the interventions used off-the-shelf video games employing the Nintendo Wii Balance Board or the Microsoft Kinect. The reasons for using VR games are its ease of access, "fun", social interaction with peers (in the case of multiplayer games), and its possible positive effect on health-related quality of life. The authors note that the studies varied greatly in terms of the interventions employed, the control conditions, the settings where the interventions were delivered and the measures of quality of life. Six of the nine studies did not demonstrate significant benefits of exergaming on quality of life. Coincidently, the three studies that did show significant effects were carried out in rehabilitation clinics and had the highest adherence rates to the program. Home-based exercise had the lowest adherence rate, what the authors argue might be due to lack of motivation to exercise alone, lack of family support, or lack of assistance from a clinician. Reasons for attrition were illness, loss of interest, lack of transportation, lack of time to complete sessions, patient deterioration and death.

They criticize the usage of small sample sizes, and the inclusion of quality of life only as a secondary outcome. The authors suggest that intensity of exergaming could be studied to find if it has any impact on quality of life. Future studies should also aim to be less heterogeneous in design and measurements used.

Martins et al. (2018) investigate modified deliveries of the Otago Exercise Program, one of which compared an AR intervention to the conventional delivery format for elderly women. AR is not defined in the study. The motivation for its use is the possibility of providing direct visual feedback to the user's on their performance. The included intervention was a screen-based system that captured the users' performance via a webcam and, based on the detected movement, sent information on the screen to guide the user. The AR version was found superior to improve balance, gait parameters and falls efficacy when compared to at-home traditional Otago Exercise Program. However, the authors criticize the study for having low methodological quality (high risk of bias).

Finally, the use of VR in sport has also been surveyed (NEUMANN et al., 2018). The authors adopt the definition that VR is a computer simulated environment that is interactive (and using interaction as an inclusion criteria) and that it induces a sense of presence. The motivation for using VR in sport is that it allows to create a controlled environment to practice and assess performance, even when coach and athletes are in different places. The sense of presence is cited as a possible mechanism for achieving these – a higher sense of presence might lead to realistic responses. In practice, the VR applications were screen or projection-based – they note that no HMDs were used. They propose a four-part framework to define VR in sport, composed of: VR Environment, sport task, athlete, and non-VR environment. All the four components result in outcomes that occur during and after the engagement in the VR sport task. These can be categorised in task performance, physiological effects, and psychological processes. They found several examples of beneficial outcomes from the use of VR, but not unanimously – which hints at other factors possibly moderating VR effectiveness (e.g. task or athlete related factors). The only outcome that had a specific measure presented was Presence, that was measured both by the ITC Sense of Presence Inventory as well as a seld-designed questionnaire. The authors criticize the homogeneous samples in the primary studies; the lack of a standardized definition of VR, sometimes confused with exergames; the lacking description of the VR applications used, and of psychological aspects of the task; and the authors suggest that Presence and Immersion measures should be standard. For future research, they point to a few paths: determining if presence changes when employing a computer screen instead of an HMD or CAVE; study the effect of the non-VR environment on performance, comparatively to the VR counterpart; study more than cycling and aerobic exercise and explore mechanical skill acquisition; study transfer to real world performance; study how VR affects psychological processes; how different factors of the VR environment affect

performance and affective outcomes; and study the use of AR.

6.9 Multiple areas

Three studies (KIM et al., 2018; SUH; PROPHET, 2018; DAVIS; NESBITT; NALIVAIKO, 2014) analyzed XR applications across multiple areas. We summarize them below.

Kim et al. (2018) reviewed the studies published on the International Symposium on Mixed and Augmented Reality (ISMAR) from 2008 to 2017. The study was carried out as a follow-up to an earlier review of the previous decade (FENG; DUH; BILLINGHURST, 2008). AR is understood as an immersive mixed environment where real and virtual things co-exist. The reviewed papers were classified into 15 non-exclusive categories, one of which was Evaluation. Among these, the study identifies 19 highly-cited papers (i.e., more than five citations per year). It divides these papers into three main categories: survey, user evaluation and perception. The use of the term "survey" by the authors meant both survey papers (i.e., that survey other research) as well as papers that report on surveys of people. Compared to the previous 10-year review of ISMAR papers, the proportion of papers related to evaluation grew from 5.8% to 15.4%. Evaluation was one of the top 5 categories both in number of publications and in number of highly cited papers.

The evaluations were further described by AR technology used (Mobile AR, see-through HMD) data collection method (user survey, user experiment, subjective survey, AR versus. non-AR experiment, experimenter subjective observation), measures (technology acceptance, experience level, performance time, placement error, user preferences, subjective feedback, accuracy, degree of realism, perceived softness, discernibility), result type (qualitative, quantitative), and domain area. Some of the evaluation papers were also present in other categories in the author's classification: Mobile-device user interface, Rendering and Visualization, Multimodal Interaction, Mid-Air Interaction, Industrial/Military Maintenance, Visualization. There are papers in the review that are not categorised as evaluation papers, but implicitly did conduct some form of evaluation, i.e., of the efficiency of tracking algorithms, the realism of rendering techniques, the ergonomics of handheld AR devices and displays, and of the effectiveness for the assessment and rehabilitation of motor dysfunction. The authors recognize this, stating there is an expectation for most ISMAR papers to include some form of evaluation.

According to the authors, work in AR evaluation going forward should focus on AR collaboration, studying AR in real world settings, what are the social, cultural and psychological phenomena behind AR, and the human perception and cognition of virtual things in AR, and devising new evaluation methods for AR. For example, new methods of evaluation could provide more accurate ways of measuring the user experience of AR systems; they point to a user-centered design and evaluation approach for VR (GABBARD; HIX; SWAN, 1999). Regarding the design of Interaction Techniques and UIs for AR, the authors argue there has been advancements on the usability of (handheld) AR. However, more research on cognitively sound interaction design is still needed for HMDs, and on the usage of physical objects as effective interaction tools.

A survey of 54 studies (SUH; PROPHET, 2018) deal with immersive technologies, a term that encompasses AR, VR and MR. Non-immersive VR studies (i.e., those that made use of 2D displays instead of an HMD) were excluded from the review. The authors place AR and VR into the reality-virtuality continuum. The main advantage of immersive technologies is the augmentation of human cognition provided by immersion – since immersion provides the "ability to perceive, feel, and cognitively process information that would have otherwise been unavailable".

They propose an adaptation of the stimulus-organism-response framework (SOR) for immersive technology. Thus, the primary applications are described in terms of sensory stimuli (i.e., displays, auditory modalities, haptic interfaces and movement tracking), perceptual stimuli (interactivity, representational fidelity, imagination, haptic imagery, perceived sense of self-location, media richness, perceived usability), and content stimuli (learning and training, psycho and physiotherapy, virtual journeys and tour, interactive simulation, gaming). Furthermore, the evaluations' outcomes are categorized in positive outcomes (learning effectiveness, learning engagement, learning attitude, task performance, reduced disease symptoms, intention to use), negative outcomes (cybersickness, physical discomfort, cognitive overload, distracted attention), as well as cognitive reactions (immersion, presence, flow, illusion, situated cognition, psychological ownership) and affective reactions (pleasure, arousal, dominance, positive/negative emotions). The data collection methods in the primary studies were categorized as case study, experiment, interview, and survey.

The authors suggest that future work should focus on evolving the constructs and measures of immersive technologies, to systematically explore how they influence the user experience and performance. For example, future research could aim to understand

how representational fidelity (or other features of immersive systems) affects learning (or other outcome of interest). They also urge for a more precise definition of immersion in general, and how it affects user performance. This encompasses understanding the mechanisms that explain how immersive environments can enhance user performance and experience. Regarding evaluation, they suggest future studies should use method triangulation to expand the currently used methods – e.g., in addition to surveys and experiments, use EEG to assess mental state and quality of user experience. They also urge researchers to take on the challenge of investigating the negative consequences of immersive technology (e.g., cybersickness and physical fatigue), and to diversify samples to more than just students.

Finally, Davis, Nesbitt and Nalivaiko (2014) review studies that evaluate cybersickness, believing it is vital to understand this phenomenon to make VR more accessible. The main goal of the review is to find whether cost-effective physiological methods to detect cybersickness during VR use exist. VR is said to be an interactive, immersive and realistic 3D simulated world – which might also be called a virtual environment. According to the authors, there is still no theory that can accurately explain the causes of cybersickness. Three potential theories are highlighted: poison theory, the postural instability theory, and sensory conflict theory. The authors identified factors that can be linked to an individual's susceptibility of experiencing cybersickness, and broke them down into three categories: Individual, Task and Device. The most popular measure of cybersickness is the SSQ, which breaks down symptoms into three, non-orthogonal factors – Nausea, Disorientation, and Oculomotor. Other subjective and objective measures of cybersickness were found. We note that one measure that aims to predict the suscepbility of cybersickness based on past experiences is not considered in our review. Some criticisms of the primary papers were their sample size, selection bias and focus on short-term results. The authors encourage future studies ensuring the accuracy of objective testing mechanisms and the cost-effectiveness analysis of physiological measures.

6.10 Industry

Only one study examined industrial applications of AR (FITE-GEORGEL, 2011). It included 54 primary studies published between 1998 and 2010. AR is defined as an environment that adds virtual elements to reality. Industrial AR (IAR) is the application of AR to support an industrial process. The authors note that these virtual elements might

be multisensory (e.g., reality could be augmented through sound rather than only visual elements) but in practice only visual augmentations were included. The main motivation for the use of AR is “aligning virtual information with the real context for the user’s benefit”.

Regarding the evaluations in the primary studies, 44% of the studies carried out some form of user testing, which were coded in three levels: no evaluation, evaluation by expert, and formal evaluation in a scientific setup to evaluate user acceptance and performance. The authors used this coding to assess the quality of the evaluations, noting that an expert review is not needed to achieve a full score. The systems that scored better in the overall quality assessment (which included the dimensions of scalability, cost-benefit and whether the system was in use out of the lab) were the ones that both performed user testing and had involvement from the industry (i.e., among the applications with the highest score, prevalence of user testing was 72%). The authors advocate for User-Centered Design in IAR system development, which encompasses the iterative evaluation of prototypes with users (i.e., formative evaluation), and formal studies when possible.

Regarding future research, the authors note that more systems could make use of the available measurements afforded by augmented tools, such as the amount of head movements, distance traveled for a picking task and voltage values in welding. Capturing these would allow for reviewing and improving worker performance as well as support the testing of new workflows. Future studies should also present cost-benefit analyses of the systems proposed. Finally, the authors believe that there is still no "reality" in IAR because only two of the primary studies had been used outside of the lab, but believe this is going to change in the near future.

7 CONCLUSIONS AND FUTURE WORK

We started this work to understand the state of XR research, which led us to investigate how XR applications are evaluated. We decided to look for reviews on XR evaluations and found out an impressive number of surveys. Due to this large landscape of survey papers, we decided to perform a tertiary review.

During the survey we also developed a set of visualizations both to support our analysis and illustrate our results.

Our study revealed which aspects of XR are most important for the domain areas that emerged from our systematic review. We identified gaps in the evaluation of XR in certain fields. Despite the large number of survey papers, there are under-represented application areas due to our inclusion criteria that targeted systematic reviews. Finally, our study allowed us to suggest a research agenda on XR evaluation.

7.1 Strengths of current XR research

The large amount of systematic literature reviews found is evidence of the widespread interest in XR across several areas. Given the year-on-year growth trend we see in the included surveys, we believe research is still accelerating. There are areas where XR has achieved encouraging evidence of its effectiveness, especially in Psychology, as shown by several meta-analyses (FERNÁNDEZ-ÁLVAREZ et al., 2019; DENG et al., 2019; CARL et al., 2019; BENBOW; ANDERSON, 2019; CARDOŞ; DAVID; DAVID, 2017; KAMPMANN; EMMELKAMP; MORINA, 2016; TURNER; CASEY, 2014), where VR exposure therapy is found to be at least as effective as in vivo exposure therapy, while potentially being easier to be accepted by patients. Learning is another area where meta-analyses show positive effects of AR (SANTOS et al., 2014; GARZÓN; ACEVEDO, 2019) and VR (MERCHANT et al., 2014) applications.

7.2 Gaps of current XR research

Comparing XR to XR is seldom done in the included studies. Only Borsci, Lawson and Broome (2015) included a primary study with XR to XR comparison, and only (CORRÊA et al., 2019) compared VR to AR. This makes it difficult to draw conclu-

sions about what specifically in XR is conducive to the outcomes seen.

Some reviews explicitly call for multi-armed factorial experiment designs to devise which characteristics of XR are responsible for the outcomes in:

- Physical Prevention (NG et al., 2019; NEUMANN et al., 2018),
- Pain Relief (CHI et al., 2019; MALLOY; MILLING, 2010),
- Post-stroke Rehabilitation (RAVI; KUMAR; SINGHI, 2017; HOWARD, 2017; VIÑAS-DIZ; SOBRIDO-PRIETO, 2016),
- Learning (BOZGEYIKLI et al., 2018; GARZÓN; ACEVEDO, 2019),
- Psychology (GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018),
- Simulators (CORRÊA et al., 2019; KENNEDY; MALDONADO; COOK, 2013), and
- Multiple areas (SUH; PROPHET, 2018).

Another gap we found is on the investigation of XR-specific and usability-related outcomes. Our analysis of the frequency of outcomes and measures found many outcomes unique to their fields (91% of outcome-measure pairs appeared in only one paper). This is expected since different fields have different domain-specific outcomes they are interested in measuring, especially when assessing the effectiveness of a new technology. However, we expected to find a hard core of XR and usability outcomes that would span all areas, but we did not. Outcomes such as Presence (n = 7), Usability (n = 9) and Cybersickness (n = 4) seldom appear in the included studies. Validated measures for them, such as the Presence Questionnaire (n = 2), System Usability Scale (n = 2), and Simulator Sickness Questionnaire (n = 2), respectively, appear even more rarely. The notable, but misleading, exceptions are the measures of Time (n = 24) and Accuracy (n = 21), which can be used to gauge Usability – these rank at the top of the most frequent measures, boosted by the papers in the Simulators category. However, for the papers in that category, they are not employed to measure Usability in a strict sense (i.e., how long it takes to perform a task with the XR application), but instead, they act as measures of the training effectiveness of the simulators. This shifts the focus of evaluation to the human rather than the artifact; instead of studying how usable an artifact is, the research is gauging if the artifact has any observable effect on the human. Another related outcome that is under-represented compared to our expectation is how ergonomic the XR application is. Using XR applications is often a whole-body experience, and human-computer interaction paradigms for XR systems are being actively studied. One main concern of HCI is how appropriate the

interaction is for the human body and cognition (i.e., physical and cognitive ergonomics). However, only 3 of our included studies cited Ergonomics as an outcome, measured via Likert scale or by collecting qualitative feedback. While it can be argued that it is the sole responsibility of the HCI field to make sure that XR applications are ergonomic, the plethora of XR hardware, software and task combinations make it unrealistic to expect the field to vet all possible forms of interaction.

We also found that some fields might be over-represented or under-represented in the landscape of surveys. Since our inclusion criteria were strict about how the surveys should be carried out (i.e., being transparent about the search strategy and inclusion criteria), some fields where systematic reviews are not as common might be under-represented in our study. On the other hand, fields where there is a long-running tradition of systematic reviews, such as Medicine and Psychology, might be over-represented. Evidence of this is that only a minority of surveys (n = 11) were found from the more technical-oriented databases (IEEE Xplore, ACM Digital), while the remaining came from ScienceDirect. An example of a research area that, in our opinion, did not get enough representation is HCI itself. While there are topics of HCI scattered across some reviews, specially (KIM et al., 2018) in the "Multiple areas" category, there are no reviews (that passed our inclusion criteria) focusing specifically on pure HCI topics, such as the evaluation of interaction techniques in XR, the evaluation of usability of XR applications, or the evaluation of the physical ergonomics of XR. An adjacent topic that did appear in our review is the study of cybersickness specifically, as in Davis, Nesbitt and Nalivaiko (2014). Finally, experiment design can be improved. A common thread among all categories is their criticism of the methodological quality of the primary studies. Even though strong designs, such as the RCT, are present in most reviews (n = 44), the authors argue that there are still several ways to improve them, such as more detailed reporting of bias, longer-term studies, and larger sample sizes.

7.3 Future research in XR

We devise three main paths of future research in XR, which we comment in this section.

Evaluating humans with XR. XR systems rely on several sensors to track users and understand the world around them. The same sensors can be used to evaluate human actions in a non-intrusive manner (such as the "System log" measures that appeared in the

reviewed papers). For example, built-in sensors can be used to objectively measure the gaze direction of the user towards something (GHIŢĂ; GUTIÉRREZ-MALDONADO, 2018), their step latency (BLUETT; BAYRAM; LITVAN, 2019), and their performance in industrial tasks (FITE-GEORGEL, 2011). During the usage of XR systems deployed to enhance human performance in a certain task, such as surgery, sensors could help objectively evaluate the surgeon's performance, not only their simulated performance as seen in the Simulators category; or they could measure a user's physical progress in stroke rehabilitation, complementing the existing scales. Moreover, the abundance of sensors may afford the measurement of task ergonomics while performing the task.

XR as a proxy to study human behavior in a safe and controlled manner. Studies focusing solely on this would not meet our inclusion criteria, since they do not involve evaluation of the XR system. However, some studies on this topic appeared as they were bundled together with other primary studies that did qualify the review to meet the inclusion criteria. In (COGNÉ et al., 2017) and (DIAS; BARBOSA; VIANNA, 2018), we see VR being used as a platform to understand human behavior, rather than to change it. In (COGNÉ et al., 2017), VR is used to understand how different environmental cues can affect human navigation; in (DIAS; BARBOSA; VIANNA, 2018), sleep-deprived surgeons performed operations on a VR simulator to study the effect of drugs on their performance. (FITE-GEORGEL, 2011) also suggests that AR sensors could be used to evaluate and optimize worker's performance and test new workflows.

Shift the focus from effectiveness to efficiency. We believe this is the main path forward for evaluation in XR. In our review, we noticed that areas (and the studies within them) are at different stages of this shift, each denoted by a main question:

1. What is XR? Using a common definition of XR. The definition of XR varied greatly from paper to paper among our included studies. For example, VR is equated to HMDs for several papers in "Pain relief", while "Physical prevention" considers video games (such as the Microsoft Kinect) VR. On the other hand, most papers in "Psychology" and "Simulators" do not define what is meant by VR or describe the systems employed in the primary studies. A common and objective taxonomy of XR (such as the three axes proposed in (MILGRAM; KISHINO, 1994)) is fundamental to achieve further stages of the efficiency shift.
2. Is XR effective? Some areas are still uncertain of the effectiveness (or lack thereof) of XR. This should be addressed by finding a common and comparable set of outcomes to be analyzed, while ensuring that the measures are valid for the outcomes.

One example is the open discussion in the usage of XR in Simulators: is the performance of students on the simulator, as recorded by it, a valid measure of the simulator's effectiveness for surgery training ? Some studies argue that actual patient outcomes are the measure that should really be studied for this outcome. Another area where XR effectiveness is not yet fully explored is Surgery (where the primary studies focused mostly on technical feasibility studies).

3. Is XR efficient? Research in this stage builds heavily on the understanding of what is XR, from the first stage, and asks the following questions:
   - Does varying the degree of one of the dimensions of XR (e.g., immersion) impact the outcomes?
   - What is the optimal "dosage" of XR?
   - Is XR cost-effective?
4. How does XR work? Research at this stage aims to understand why XR is effective/efficient and is divided into two types of research:
   - What is the *general* human response to XR: building foundational knowledge on how the human brain and body are affected by XR.
   - What is the *specific* human response to XR: applying this general knowledge and devising specific rules and exceptions for different tasks and applications.

7.4 Future research in survey visualizations

Current methods for visualizing surveys have a main limitation when applied to our dataset: the attributes extracted from the studies are linked directly to the publication, and as a consequence, it is cumbersome and sometimes impossible to represent attributes that are part of a network.

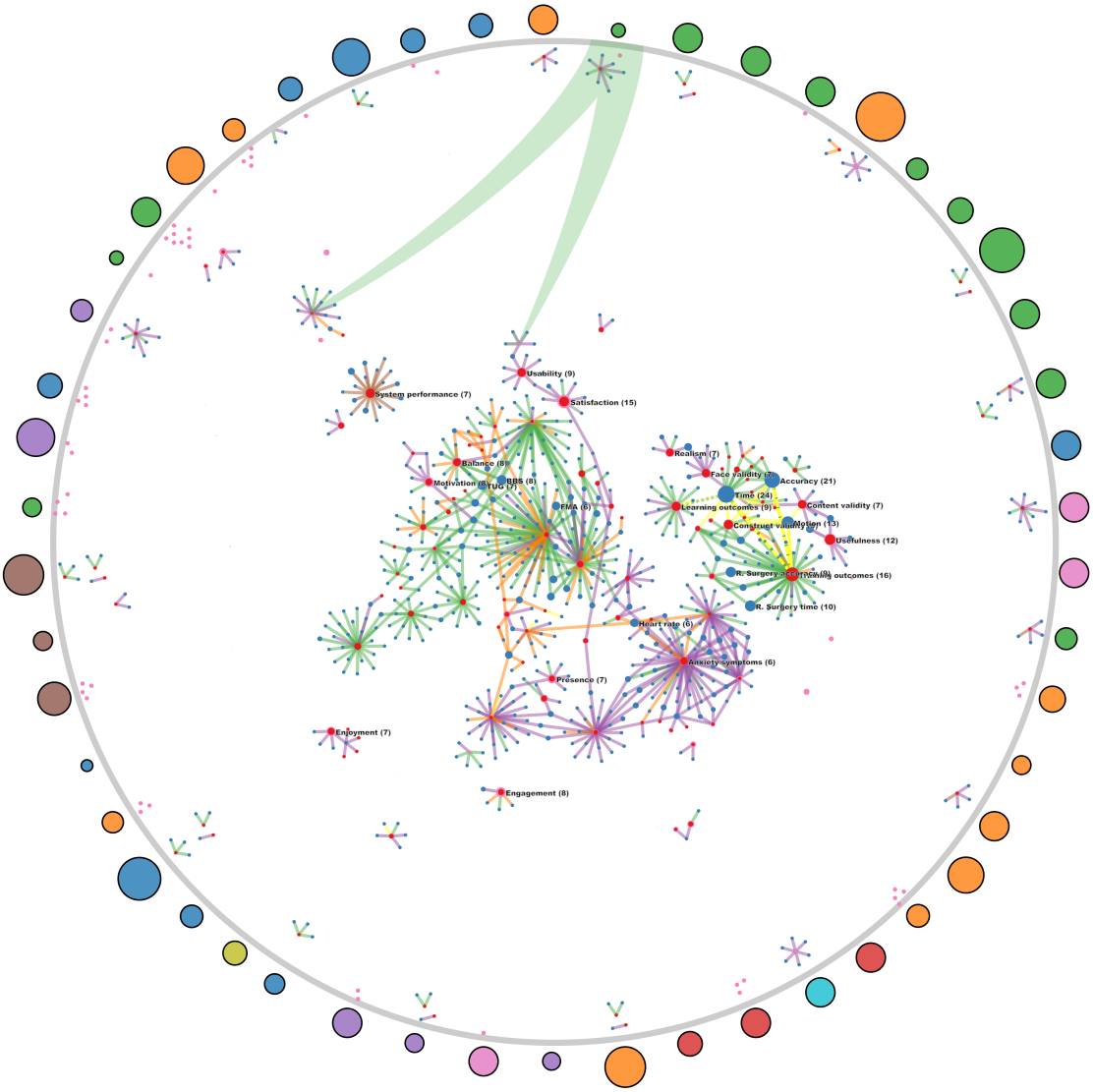
We believe that even secondary reviews might have attributes that form networks. For example, if we repeated our study as a secondary review, we would still extract an evaluation network of outcomes and measures from each primary study. Thus, advances in layered network visualization could be useful for both secondary and tertiary reviews. The biggest difficulty we found when designing a visualization was connecting the publication layer with the attribute layers. While currently quilts can nearly solve this problem, they come with the issues that are common to all tabular layouts: it is hard to present more

than 100 data points at a time, visualizing paths is harder than with other layouts, and the network topology can be more or less evident depending on the solution used for row and column ordering. We think there are still advances to be made in the visualization of layered networks, especially those that are not purely hierarchical (and could be visualized with containment marks, with techniques such as GrouseFlocks (ARCHAMBAULT; MUNZNER; AUBER, 2008)).

One potential path forward would be to integrate all layers in a single visualization, making it much more expressive in terms of the shared topologies of publications – seeing what graphs occur only on a single paper and what graphs are shared across papers. While there are ways to achieve this, such as simply including the publications on the node-link diagram or using quilts – both with some drawbacks, as noted in Sec. 4, we think a potential novel solution would be to incorporate a radial network visualization for the top level (the publications). Since the publications have no links to one another in our dataset, the links would only go to the attributes. This would result in a potentially less cluttered network visualization than a pure node-link diagram, since attributes that are present in only one publication could be clumped together near the publication – even excusing them from needing to have link marks to it. The problem that remains is showing the path across the two attribute levels unambiguously. This would require some more complex link marks, or the use of interaction.

Figure 7.1 shows a mock-up rendition of such visualization: the publications are arranged in an outer ring, and hovering over one highlights the area of its exclusive graph, which smoothly turns into links for the shared graphs – augmenting the links that already exist to disambiguate what measures are contained by that publication.

Figure 7.1: Mock-up of a visualization to represent layered networks, where the top layer is arranged in a radial layout, and the other layers in a graph inside that layout.



REFERENCES

AHMED, K. et al. Effectiveness of Procedural Simulation in Urology: A Systematic Review. **Journal of Urology**, v. 186, n. 1, p. 26–34, jul. 2011. Publisher: WoltersKluwer. Available from Internet: <https://www.auajournals.org/doi/10.1016/j.juro.2011.02.2684>.

AÏM, F. et al. Effectiveness of Virtual Reality Training in Orthopaedic Surgery. **Arthroscopy: The Journal of Arthroscopic & Related Surgery**, v. 32, n. 1, p. 224–232, jan. 2016. ISSN 0749-8063. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0749806315006489>.

AKÇAYIR, M.; AKÇAYIR, G. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. **Educational Research Review**, v. 20, p. 1–11, feb. 2017. ISSN 1747-938X. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1747938X16300616>.

ALAKER, M.; WYNN, G. R.; ARULAMPALAM, T. Virtual reality training in laparoscopic surgery: A systematic review & meta-analysis. **International Journal of Surgery**, v. 29, p. 85–94, may 2016. ISSN 1743-9191. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S174391911600251X>.

ALASHRAM, A. R. et al. Cognitive rehabilitation post traumatic brain injury: A systematic review for emerging use of virtual reality technology. **Journal of Clinical Neuroscience**, v. 66, p. 209–219, aug. 2019. ISSN 0967-5868. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0967586819305600>.

ANDERSON-HANLEY, C.; ARCIERO, P.; Snyder. Social facilitation in virtual reality-enhanced exercise: competitiveness moderates exercise effort of older adults. **Clinical Interventions in Aging**, p. 275, oct. 2011. ISSN 1178-1998. Available from Internet: <http://www.dovepress.com/social-facilitation-in-virtual-reality-enhanced-exercise-competitivene-peer-reviewed-article-CIA>.

ARCHAMBAULT, D.; MUNZNER, T.; AUBER, D. GrouseFlocks: Steerable Exploration of Graph Hierarchy Space. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 4, p. 900–913, jul. 2008. ISSN 1077-2626. Available from Internet: <http://ieeexplore.ieee.org/document/4447668/>.

ARORA, A. et al. Virtual reality simulation training in Otolaryngology. **International Journal of Surgery**, v. 12, n. 2, p. 87–94, feb. 2014. ISSN 1743-9191. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1743919113011072>.

AZUMA, R. et al. Recent advances in augmented reality. **IEEE Computer Graphics and Applications**, v. 21, n. 6, p. 34–47, dec. 2001. ISSN 02721716. Available from Internet: <http://ieeexplore.ieee.org/document/963459/>.

AZUMA, R. T. A Survey of Augmented Reality. **Presence: Teleoperators and Virtual Environments**, v. 6, n. 4, p. 355–385, aug. 1997. ISSN 1054-7460. Available from Internet: <https://direct.mit.edu/pvar/article/6/4/355-385/18336>.

BAÑOS, R. et al. Presence and Reality Judgment in Virtual Environments: A Unitary Construct? **CyberPsychology & Behavior**, v. 3, n. 3, p. 327–335, jun. 2000. ISSN 1094-9313, 1557-8364. Available from Internet: <http://www.liebertpub.com/doi/10.1089/10949310050078760>.

BAÑOS, R. et al. A virtual reality system for the treatment of stress-related disorders: A preliminary analysis of efficacy compared to a standard cognitive behavioral program. **International Journal of Human-Computer Studies**, v. 69, n. 9, p. 602–613, aug. 2011. ISSN 10715819. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S1071581911000656>.

BECK, F. et al. The State of the Art in Visualizing Dynamic Graphs. In: **EuroVis**. [S.l.: s.n.], 2014.

BECK, F.; KOCH, S.; WEISKOPF, D. Visual Analysis and Dissemination of Scientific Literature Collections with SurVis. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 180–189, jan. 2016. ISSN 1941-0506. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

BECK, F.; WEISKOPF, D. Word-Sized Graphics for Scientific Texts. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 6, p. 1576–1587, jun. 2017. ISSN 1941-0506. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

BENBOW, A. A.; ANDERSON, P. L. A meta-analytic examination of attrition in virtual reality exposure therapy for anxiety disorders. **Journal of Anxiety Disorders**, v. 61, p. 18–26, jan. 2019. ISSN 0887-6185. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0887618518300525>.

BERNHARDT, S. et al. The status of augmented reality in laparoscopic surgery as of 2016. **Medical Image Analysis**, v. 37, p. 66–90, abr. 2017. ISSN 1361-8415. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1361841517300178>.

BLATTGERSTE, J.; RENNER, P.; PFEIFFER, T. Augmented reality action assistance and learning for cognitively impaired people: a systematic literature review. In: **Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments**. Rhodes, Greece: Association for Computing Machinery, 2019. (PETRA '19), p. 270–279. ISBN 978-1-4503-6232-0. Available from Internet: <https://doi.org/10.1145/3316782.3316789>.

BLUETT, B.; BAYRAM, E.; LITVAN, I. The virtual reality of Parkinson's disease freezing of gait: A systematic review. **Parkinsonism & Related Disorders**, v. 61, p. 26–33, abr. 2019. ISSN 1353-8020. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S135380201830498X>.

BOAS, Y. A. G. V. **Overview of Virtual Reality Technologies**. 2012. Available from Internet: </paper/Overview-of-Virtual-Reality-Technologies-Boas/4214cb09e29795f5363e5e3b545750dce027b668>.

BORSCI, S.; LAWSON, G.; BROOME, S. Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance. **Computers in Industry**,

v. 67, p. 17–26, feb. 2015. ISSN 0166-3615. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0166361514002073>.

BOSC, R. et al. Intraoperative augmented reality with heads-up displays in maxillofacial surgery: a systematic review of the literature and a classification of relevant technologies. **International Journal of Oral and Maxillofacial Surgery**, v. 48, n. 1, p. 132–139, jan. 2019. ISSN 0901-5027. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0901502718303801>.

BOTELLA, C. et al. Virtual reality exposure in the treatment of panic disorder and agoraphobia: A controlled study. **Clinical Psychology & Psychotherapy**, v. 14, n. 3, p. 164–175, may 2007. ISSN 10633995, 10990879. Available from Internet: <http://doi.wiley.com/10.1002/cpp.524>.

BOWMAN, D. A. (Ed.). **3D user interfaces: theory and practice**. Boston: Addison-Wesley, 2005. ISBN 978-0-201-75867-2.

BOWMAN, D. A.; GABBARD, J. L.; HIX, D. A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. **Presence: Teleoperators and Virtual Environments**, v. 11, n. 4, p. 404–424, aug. 2002. ISSN 1054-7460. Available from Internet: <https://direct.mit.edu/pvar/article/11/4/404-424/18421>.

BOZGEYIKLI, L. et al. A Survey on Virtual Reality for Individuals with Autism Spectrum Disorder: Design Considerations. **IEEE Transactions on Learning Technologies**, v. 11, n. 2, p. 133–151, abr. 2018. ISSN 1939-1382, 2372-0050. Available from Internet: <https://ieeexplore.ieee.org/document/8010470/>.

BRUNCKHORST, O. et al. Simulation-Based Ureteroscopy Training: A Systematic Review. **Journal of Surgical Education**, v. 72, n. 1, p. 135–143, jan. 2015. ISSN 1931-7204. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1931720414002037>.

BUDGEN, D. et al. Reporting systematic reviews: Some lessons from a tertiary study. **Information and Software Technology**, v. 95, p. 62–74, mar. 2018. ISSN 09505849. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0950584916303548>.

BURDEA, G.; COIFFET, P. Virtual Reality Technology. **Presence: Teleoperators and Virtual Environments**, v. 12, n. 6, p. 663–664, dec. 2003. ISSN 1054-7460. Available from Internet: <https://direct.mit.edu/pvar/article/12/6/663-664/18476>.

CACCIATA, M. et al. Effect of exergaming on health-related quality of life in older adults: A systematic review. **International Journal of Nursing Studies**, v. 93, p. 30–40, may 2019. ISSN 0020-7489. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0020748919300239>.

CALDERÕN, M. A. F. et al. Analysis of the Factors Related to the Effectiveness of Transcranial Current Stimulation in Upper Limb Motor Function Recovery after Stroke: a Systematic Review. **Journal of Medical Systems**, v. 43, n. 3, p. 69, feb. 2019. ISSN 1573-689X. Available from Internet: <https://doi.org/10.1007/s10916-019-1193-9>.

CARD, S.; MACKINLAY, J.; SHNEIDERMAN, B. **Readings in information visualization: using vision to think.** [S.l.]: Morgan Kaufmann, 1999.

CÁRDENAS, K.; ARANDA, M. Psychotherapies for the treatment of phantom limb pain. **Revista Colombiana de Psiquiatría (English ed.)**, v. 46, n. 3, p. 178–186, jul. 2017. ISSN 2530-3120. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S2530312017300450>.

CARDOŞ, R. A. I.; DAVID, O. A.; DAVID, D. O. Virtual reality exposure therapy in flight anxiety: A quantitative meta-analysis. **Computers in Human Behavior**, v. 72, p. 371–380, jul. 2017. ISSN 0747-5632. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0747563217301565>.

CARL, E. et al. Virtual reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials. **Journal of Anxiety Disorders**, v. 61, p. 27–36, jan. 2019. ISSN 0887-6185. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0887618518302469>.

CARMIGNIANI, J. et al. Augmented reality technologies, systems and applications. **Multimedia Tools and Applications**, v. 51, n. 1, p. 341–377, jan. 2011. ISSN 1380-7501, 1573-7721. Available from Internet: <http://link.springer.com/10.1007/s11042-010-0660-6>.

CAUDELL, T.; MIZELL, D. Augmented reality: an application of heads-up display technology to manual manufacturing processes. In: **Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences**. Kauai, HI, USA: IEEE, 1992. p. 659–669 vol.2. ISBN 978-0-8186-2420-9. Available from Internet: <http://ieeexplore.ieee.org/document/183317/>.

CHATZIMPARMPAS, A. et al. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. 2020. ISSN 1467-8659. Accepted: 2020-05-24T13:53:52Z Publisher: The Eurographics Association and John Wiley & Sons Ltd. Available from Internet: <https://diglib.eg.org:443/xmlui/handle/10.1111/cgf14034>.

CHATZIMPARMPAS, A. et al. A survey of surveys on the use of visualization for interpreting machine learning models. **Information Visualization**, v. 19, n. 3, p. 207–233, jul. 2020. ISSN 1473-8716. Publisher: SAGE Publications. Available from Internet: <https://doi.org/10.1177/1473871620904671>.

CHEN, Y. et al. Home-based technologies for stroke rehabilitation: A systematic review. **International Journal of Medical Informatics**, v. 123, p. 11–22, mar. 2019. ISSN 1386-5056. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1386505618302740>.

CHI, B. et al. Virtual reality for spinal cord injury-associated neuropathic pain: Systematic review. **Annals of Physical and Rehabilitation Medicine**, v. 62, n. 1, p. 49–57, jan. 2019. ISSN 1877-0657. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1877065718314532>.

CLARK, A. D. et al. The Effect of 3-Dimensional Simulation on Neurosurgical Skill Acquisition and Surgical Performance: A Review of the Literature. **Journal of Surgical**

Education, v. 74, n. 5, p. 828–836, sep. 2017. ISSN 1931-7204. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1931720416303166>.

COGNÉ, M. et al. The contribution of virtual reality to the diagnosis of spatial navigation disorders and to the study of the role of navigational aids: A systematic literature review. **Annals of Physical and Rehabilitation Medicine**, v. 60, n. 3, p. 164–176, jun. 2017. ISSN 1877-0657. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1877065716000026>.

CORBETTA, D.; IMERI, F.; GATTI, R. Rehabilitation that incorporates virtual reality is more effective than standard rehabilitation for improving walking speed, balance and mobility after stroke: a systematic review. **Journal of Physiotherapy**, v. 61, n. 3, p. 117–124, jul. 2015. ISSN 1836-9553. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1836955315000569>.

CORRÊA, C. G. et al. Haptic interaction for needle insertion training in medical applications: The state-of-the-art. **Medical Engineering & Physics**, v. 63, p. 6–25, jan. 2019. ISSN 1350-4533. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S135045331830167X>.

COYLE, H.; TRAYNOR, V.; SOLOWIJ, N. Computerized and Virtual Reality Cognitive Training for Individuals at High Risk of Cognitive Decline: Systematic Review of the Literature. **The American Journal of Geriatric Psychiatry**, v. 23, n. 4, p. 335–359, abr. 2015. ISSN 1064-7481. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1064748114001390>.

CRUZ-NEIRA, C.; SANDIN, D. J.; DEFANTI, T. A. Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In: **Proceedings of the 20th annual conference on Computer graphics and interactive techniques - SIGGRAPH '93**. Not Known: ACM Press, 1993. p. 135–142. ISBN 978-0-89791-601-1. Available from Internet: <http://portal.acm.org/citation.cfm?doid=166117.166134>.

DAVIS, S.; NESBITT, K.; NALIVAIKO, E. A Systematic Review of Cybersickness. In: **Proceedings of the 2014 Conference on Interactive Entertainment**. New York, NY, USA: Association for Computing Machinery, 2014. (IE2014), p. 1–9. ISBN 978-1-4503-2790-9. Available from Internet: <https://doi.org/10.1145/2677758.2677780>.

DENG, W. et al. The efficacy of virtual reality exposure therapy for PTSD symptoms: A systematic review and meta-analysis. **Journal of Affective Disorders**, v. 257, p. 698–709, oct. 2019. ISSN 0165-0327. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0165032719308328>.

DETMER, F. J. et al. Virtual and Augmented Reality Systems for Renal Interventions: A Systematic Review. **IEEE Reviews in Biomedical Engineering**, v. 10, p. 78–94, 2017. ISSN 1937-3333, 1941-1189. Available from Internet: <http://ieeexplore.ieee.org/document/8026164/>.

DIAS, L. P. S.; BARBOSA, J. L. V.; VIANNA, H. D. Gamification and serious games in depression care: A systematic mapping study. **Telematics and Informatics**, v. 35, n. 1, p. 213–224, abr. 2018. ISSN 0736-5853. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0736585317305865>.

DUMAS, M.; MCGUFFIN, M. **Financevis . netA Visual Survey of Financial Data Visualizations**. 2014. Available from Internet: <https://www.semanticscholar.org/paper/Financevis-.-netA-Visual-Survey-of-Financial-Data-Dumas-McGuffin/13415197345fba3f11683ac9f860becabfa78123>.

DUNLEAVY, M.; DEDE, C.; MITCHELL, R. Affordances and Limitations of Immersive Participatory Augmented Reality Simulations for Teaching and Learning. **Journal of Science Education and Technology**, v. 18, n. 1, p. 7–22, feb. 2009. ISSN 1059-0145, 1573-1839. Available from Internet: <http://link.springer.com/10.1007/s10956-008-9119-1>.

ECK, N. J. van; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, v. 84, n. 2, p. 523–538, aug. 2010. ISSN 1588-2861. Available from Internet: <https://doi.org/10.1007/s11192-009-0146-3>.

EITAN, A. et al. **Connected Papers | Find and explore academic papers**. 2021. Available from Internet: <https://www.connectedpapers.com/>.

ELKIND, J. S. et al. A Simulated Reality Scenario Compared with the Computerized Wisconsin Card Sorting Test: An Analysis of Preliminary Results. **CyberPsychology & Behavior**, v. 4, n. 4, p. 489–496, aug. 2001. ISSN 1094-9313, 1557-8364. Available from Internet: <http://www.liebertpub.com/doi/10.1089/109493101750527042>.

EXPERIENCE, W. L. i. R.-B. U. **Usability 101: Introduction to Usability**. 2012. Available from Internet: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>.

FABBRI, S. et al. Managing Literature Reviews Information through Visualization:. In: **Proceedings of the 14th International Conference on Enterprise Information Systems**. Wroclaw, Poland: SciTePress - Science and and Technology Publications, 2012. p. 36–45. ISBN 978-989-8565-10-5 978-989-8565-11-2 978-989-8565-12-9. Available from Internet: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0004004000360045>.

FEDERICO, P. et al. A Survey on Visual Approaches for Analyzing Scientific Literature and Patents. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 9, p. 2179–2198, sep. 2017. ISSN 1941-0506. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

FEINER, S. K. Augmented Reality: A New Way of Seeing. **Scientific American**, v. 286, n. 4, p. 48–55, abr. 2002. ISSN 0036-8733. Available from Internet: <https://www.scientificamerican.com/article/augmented-reality-a-new-w>.

FENG, Z.; DUH, H. B.-L.; BILLINGHURST, M. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In: **2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality**. Cambridge, UK: IEEE, 2008. p. 193–202. ISBN 978-1-4244-2840-3. Available from Internet: <http://ieeexplore.ieee.org/document/4637362/>.

FENG, Z. et al. Immersive virtual reality serious games for evacuation training and research: A systematic literature review. **Computers & Education**, v. 127, p. 252–266,

dec. 2018. ISSN 0360-1315. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0360131518302380>.

FERNÁNDEZ-ÁLVAREZ, J. et al. Deterioration rates in Virtual Reality Therapy: An individual patient data level meta-analysis. **Journal of Anxiety Disorders**, v. 61, p. 3–17, jan. 2019. ISSN 0887-6185. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0887618517306114>.

FITE-GEORGEL, P. Is there a reality in Industrial Augmented Reality? In: **2011 10th IEEE International Symposium on Mixed and Augmented Reality**. Basel: IEEE, 2011. p. 201–210. ISBN 978-1-4577-2185-4 978-1-4577-2183-0 978-1-4577-2184-7. Available from Internet: <http://ieeexplore.ieee.org/document/6162889/>.

FOX, J.; ARENA, D.; BAILENSON, J. N. Virtual Reality: A Survival Guide for the Social Scientist. **Journal of Media Psychology**, v. 21, n. 3, p. 95–113, jan. 2009. ISSN 1864-1105, 2151-2388. Available from Internet: <https://econtent.hogrefe.com/doi/10.1027/1864-1105.21.3.95>.

GABBARD, J.; HIX, D.; SWAN, J. User-centered design and evaluation of virtual environments. **IEEE Computer Graphics and Applications**, v. 19, n. 6, p. 51–59, nov. 1999. ISSN 1558-1756. Conference Name: IEEE Computer Graphics and Applications.

GAMBERINI, L. Virtual Reality as a New Research Tool for the Study of Human Memory. **CyberPsychology & Behavior**, v. 3, n. 3, p. 337–342, jun. 2000. ISSN 1094-9313, 1557-8364. Available from Internet: <http://www.liebertpub.com/doi/10.1089/10949310050078779>.

GAO, Y.; GONZALEZ, V. A.; YIU, T. W. The effectiveness of traditional tools and computer-aided technologies for health and safety training in the construction sector: A systematic review. **Computers & Education**, v. 138, p. 101–115, sep. 2019. ISSN 0360-1315. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0360131519301125>.

GARZÓN, J.; ACEVEDO, J. Meta-analysis of the impact of Augmented Reality on students' learning gains. **Educational Research Review**, v. 27, p. 244–260, jun. 2019. ISSN 1747-938X. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1747938X18301805>.

GERARDI, M. et al. Virtual Reality Exposure Therapy for Post-Traumatic Stress Disorder and Other Anxiety Disorders. **Current Psychiatry Reports**, v. 12, n. 4, p. 298–305, aug. 2010. ISSN 1523-3812, 1535-1645. Available from Internet: <http://link.springer.com/10.1007/s11920-010-0128-4>.

GERARDI, M. et al. Virtual reality exposure therapy using a virtual Iraq: Case report. **Journal of Traumatic Stress**, v. 21, n. 2, p. 209–213, abr. 2008. ISSN 08949867, 15736598. Available from Internet: <http://doi.wiley.com/10.1002/jts.20331>.

GHIŢĂ, A.; GUTIÉRREZ-MALDONADO, J. Applications of virtual reality in individuals with alcohol misuse: A systematic review. **Addictive Behaviors**, v. 81, p. 1–11, jun. 2018. ISSN 0306-4603. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0306460318300492>.

GRECO, F. et al. Current Perspectives in the Use of Molecular Imaging To Target Surgical Treatments for Genitourinary Cancers. **European Urology**, v. 65, n. 5, p. 947–964, may 2014. ISSN 0302-2838. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0302283813007483>.

GUEDES, H. G. et al. Virtual reality simulator versus box-trainer to teach minimally invasive procedures: A meta-analysis. **International Journal of Surgery**, v. 61, p. 60–68, jan. 2019. ISSN 1743-9191. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1743919118317229>.

GUJJAR, K. R. et al. Are Technology-Based Interventions Effective in Reducing Dental Anxiety in Children and Adults? A Systematic Review. **Journal of Evidence Based Dental Practice**, v. 19, n. 2, p. 140–155, jun. 2019. ISSN 1532-3382. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1532338218302896>.

GUO, H.; YU, Y.; SKITMORE, M. Visualization technology-based construction safety management: A review. **Automation in Construction**, v. 73, p. 135–144, jan. 2017. ISSN 09265805. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S092658051630293X>.

HALE, K. S.; STANNEY, K. M. (Ed.). **Handbook of virtual environments: design, implementation, and applications**. Second edition. Boca Raton: CRC Press, Taylor & Francis Group, 2015. (Human factors and ergonomics). ISBN 978-1-4665-1184-2.

HALLER, M.; BILLINGHURST, M.; THOMAS, B. (Ed.). **Emerging Technologies of Augmented Reality: Interfaces and Design**. IGI Global, 2007. ISBN 978-1-59904-066-0 978-1-59904-068-4. Available from Internet: <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59904-066-0>.

HAN, J. et al. Mapping the intellectual structure of research on surgery with mixed reality: Bibliometric network analysis (2000–2019). **Journal of Biomedical Informatics**, v. 109, p. 103516, sep. 2020. ISSN 1532-0464. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1532046420301441>.

HINCKLEY, K. Input Technologies and Techniques. jan. 2002. ISSN 978-1-4200-8881-6.

HOFFMAN, H. G. et al. Effectiveness of Virtual Reality–Based Pain Control With Multiple Treatments:. **The Clinical Journal of Pain**, v. 17, n. 3, p. 229–235, sep. 2001. ISSN 0749-8047. Available from Internet: <http://journals.lww.com/00002508-200109000-00007>.

HOWARD, M. C. A meta-analysis and systematic literature review of virtual reality rehabilitation programs. **Computers in Human Behavior**, v. 70, p. 317–327, may 2017. ISSN 0747-5632. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0747563217300134>.

HUANG, Y. et al. Rehabilitation using virtual reality technology: a bibliometric analysis, 1996–2015. **Scientometrics**, v. 109, n. 3, p. 1547–1559, dec. 2016. ISSN 1588-2861. Available from Internet: <https://doi.org/10.1007/s11192-016-2117-9>.

IBÁÑEZ, M.-B.; DELGADO-KLOOS, C. Augmented reality for STEM learning: A systematic review. **Computers & Education**, v. 123, p. 109–123, aug. 2018. ISSN 0360-1315. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0360131518301027>.

ISO. **ISO 9241-11:2018**. 2018. Available from Internet: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/35/63500.html>.

ISO/IEC. **ISO/IEC 25066:2016**. 2016. Available from Internet: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/38/63831.html>.

JODA, T. et al. Augmented and virtual reality in dental medicine: A systematic review. **Computers in Biology and Medicine**, v. 108, p. 93–100, may 2019. ISSN 0010-4825. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S001048251930085X>.

KAMPMANN, I. L.; EMMELKAMP, P. M. G.; MORINA, N. Meta-analysis of technology-assisted interventions for social anxiety disorder. **Journal of Anxiety Disorders**, v. 42, p. 71–84, aug. 2016. ISSN 0887-6185. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0887618516300974>.

KARAKUS, M.; ERSOZLU, A.; CLARK, A. C. Augmented Reality Research in Education: A Bibliometric Study. **Eurasia Journal of Mathematics, Science and Technology Education**, v. 15, n. 10, p. em1755, may 2019. ISSN 1305-8215, 1305-8223. Publisher: Modestum Publishing LTD. Available from Internet: <https://www.ejmste.com/article/augmented-reality-research-in-education-a-bibliometric-study-7712>.

KEHRER, J.; HAUSER, H. Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 3, p. 495–513, mar. 2013. ISSN 1941-0506. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

KENNEDY, C. C.; MALDONADO, F.; COOK, D. A. Simulation-Based Bronchoscopy Training: Systematic Review and Meta-analysis. **Chest**, v. 144, n. 1, p. 183–192, jul. 2013. ISSN 0012-3692. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0012369213604663>.

KERREN, A. et al. BioVis Explorer: A visual guide for biological data visualization techniques. **PLOS ONE**, v. 12, n. 11, p. e0187341, nov. 2017. ISSN 1932-6203. Publisher: Public Library of Science. Available from Internet: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187341>.

KERSTEN-OERTEL, M.; JANNIN, P.; COLLINS, D. L. The state of the art of visualization in mixed reality image guided surgery. **Computerized Medical Imaging and Graphics**, v. 37, n. 2, p. 98–112, mar. 2013. ISSN 0895-6111. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0895611113000116>.

KIM, K. et al. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 11, p. 2947–2962, nov. 2018. ISSN 1077-2626, 1941-0506, 2160-9306. Available from Internet: <https://ieeexplore.ieee.org/document/8456568/>.

KITCHENHAM, B. **Procedures for Performing Systematic Reviews**. [S.l.], 2004. 34 p.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for Performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering – A tertiary study. **Information and Software Technology**, v. 52, n. 8, p. 792–805, aug. 2010. ISSN 09505849. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0950584910000467>.

KLINKE, M. E. et al. Ward-based interventions for patients with hemispatial neglect in stroke rehabilitation: A systematic literature review. **International Journal of Nursing Studies**, v. 52, n. 8, p. 1375–1403, aug. 2015. ISSN 0020-7489. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0020748915001054>.

KNIGHT, R. G.; TITOV, N. Use of Virtual Reality Tasks to Assess Prospective Memory: Applicability and Evidence. **Brain Impairment**, v. 10, n. 1, p. 3–13, may 2009. ISSN 1443-9646, 1839-5252. Available from Internet: <https://www.cambridge.org/core/product/identifier/S1443964600001728/type/journal_article>.

KOLASINSKI, E. M. **Simulator Sickness in Virtual Environments**. [S.l.], 1995. 68 p.

KOSKINA, A.; CAMPBELL, I. C.; SCHMIDT, U. Exposure therapy in eating disorders revisited. **Neuroscience & Biobehavioral Reviews**, v. 37, n. 2, p. 193–208, feb. 2013. ISSN 0149-7634. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S014976341200200X>.

KU, J. et al. A Virtual Environment for Investigating Schizophrenic Patients' Characteristics: Assessment of Cognitive and Navigation Ability. **CyberPsychology & Behavior**, v. 6, n. 4, p. 397–404, aug. 2003. ISSN 1094-9313, 1557-8364. Available from Internet: <http://www.liebertpub.com/doi/10.1089/109493103322278781>.

KUCHER, K.; KERREN, A. Text visualization techniques: Taxonomy, visual survey, and community insights. In: **2015 IEEE Pacific Visualization Symposium (PacificVis)**. [S.l.: s.n.], 2015. p. 117–121. ISSN: 2165-8773.

KUCHER, K.; PARADIS, C.; KERREN, A. The State of the Art in Sentiment Visualization. **Computer Graphics Forum**, v. 37, n. 1, p. 71–96, 2018. ISSN 1467-8659. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13217. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13217>.

KURILOVAS, E. Evaluation of quality and personalisation of VR/AR/MR learning systems. **Behaviour & Information Technology**, v. 35, n. 11, p. 998–1007, nov. 2016. ISSN 0144-929X. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2016.1212929. Available from Internet: <https://doi.org/10.1080/0144929X.2016.1212929>.

LACERDA, G. et al. Code smells and refactoring: A tertiary systematic review of challenges and observations. **Journal of Systems and Software**, v. 167, p. 110610, sep. 2020. ISSN 0164-1212. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0164121220300881>.

LALONDE, G. et al. Assessment of executive function in adolescence: A comparison of traditional and virtual reality tools. **Journal of Neuroscience Methods**, v. 219, n. 1, p. 76–82, sep. 2013. ISSN 01650270. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0165027013002434>.

LAVALLE, S. **Virtual Reality - LaValle**. 2015. Available from Internet: <http://lavalle.pl/vr/>.

LAVER, K. et al. Virtual reality stroke rehabilitation - hype or hope?: VIEWPOINT. **Australian Occupational Therapy Journal**, v. 58, n. 3, p. 215–219, jun. 2011. ISSN 00450766. Available from Internet: <http://doi.wiley.com/10.1111/j.1440-1630.2010.00897.x>.

LAVIOLA, J. J. A discussion of cybersickness in virtual environments. **ACM SIGCHI Bulletin**, v. 32, n. 1, p. 47–56, jan. 2000. ISSN 0736-6906. Available from Internet: <https://dl.acm.org/doi/10.1145/333329.333344>.

LAVIOLA, J. J. et al. **3D user interfaces: theory and practice**. Second edition. Boston: Addison-Wesley, 2017. (Addison-Wesley usability and HCI series). OCLC: ocn935986831. ISBN 978-0-13-403432-4.

LEE, H.-G.; CHUNG, S.; LEE, W.-H. Presence in virtual golf simulators: The effects of presence on perceived enjoyment, perceived value, and behavioral intention. **New Media & Society**, v. 15, n. 6, p. 930–946, sep. 2013. ISSN 1461-4448, 1461-7315. Available from Internet: <http://journals.sagepub.com/doi/10.1177/1461444812464033>.

LIN, I.-H. et al. Effectiveness and Superiority of Rehabilitative Treatments in Enhancing Motor Recovery Within 6 Months Poststroke: A Systemic Review. **Archives of Physical Medicine and Rehabilitation**, v. 100, n. 2, p. 366–378, feb. 2019. ISSN 0003-9993. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0003999318313972>.

LIU, S. et al. Visualizing High-Dimensional Data: Advances in the Past Decade. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 3, p. 1249–1268, mar. 2017. ISSN 1941-0506. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

LOWOOD, H. **Virtual reality**. 2015. Available from Internet: <https://www.britannica.com/technology/virtual-reality>.

LU, Y. et al. The State-of-the-Art in Predictive Visual Analytics. **Computer Graphics Forum**, v. 36, n. 3, p. 539–562, 2017. ISSN 1467-8659. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13210. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13210>.

MALLOY, K. M.; MILLING, L. S. The effectiveness of virtual reality distraction for pain reduction: A systematic review. **Clinical Psychology Review**, v. 30, n. 8, p. 1011–1018, dec. 2010. ISSN 0272-7358. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0272735810001091>.

MARTIN, S. et al. Effectiveness and impact of networked communication interventions in young people with mental health conditions: A systematic review. **Patient Education and Counseling**, v. 85, n. 2, p. e108–e119, nov. 2011. ISSN 0738-3991. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0738399110006993>.

MARTINS, A. C. et al. Does modified Otago Exercise Program improves balance in older people? A systematic review. **Preventive Medicine Reports**, v. 11, p. 231–239, sep. 2018. ISSN 2211-3355. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S2211335518301116>.

MAZUR, T. et al. Virtual Reality–Based Simulators for Cranial Tumor Surgery: A Systematic Review. **World Neurosurgery**, v. 110, p. 414–422, feb. 2018. ISSN 1878-8750. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1878875017320648>.

MCCLOY, R.; STONE, R. Science, medicine, and the future: Virtual reality in surgery. **BMJ**, v. 323, n. 7318, p. 912–915, oct. 2001. ISSN 0959-8138, 1468-5833. Available from Internet: <https://www.bmj.com/lookup/doi/10.1136/bmj.323.7318.912>.

MCNABB, L.; LARAMEE, R. S. Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization. **Computer Graphics Forum**, v. 36, n. 3, p. 589–617, jun. 2017. ISSN 0167-7055, 1467-8659. Available from Internet: <https://onlinelibrary.wiley.com/doi/10.1111/cgf.13212>.

MERCHANT, Z. et al. Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. **Computers & Education**, v. 70, p. 29–40, jan. 2014. ISSN 0360-1315. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0360131513002108>.

MERIANS, A. S. et al. Virtual Reality–Augmented Rehabilitation for Patients Following Stroke. **Physical Therapy**, v. 82, n. 9, p. 898–915, sep. 2002. ISSN 0031-9023, 1538-6724. Available from Internet: <https://academic.oup.com/ptj/article/82/9/898/2857676>.

MILGRAM, P.; KISHINO, F. A Taxonomy of Mixed Reality Visual Displays. **undefined**, 1994. Available from Internet: </paper/A-Taxonomy-of-Mixed-Reality-Visual-Displays-Milgram-Kishino/f78a31be8874eda176a5244c645289be9f1d4317>.

MILLER, H. L.; BUGNARIU, N. L. Level of Immersion in Virtual Environments Impacts the Ability to Assess and Teach Social Skills in Autism Spectrum Disorder. **Cyberpsychology, Behavior, and Social Networking**, v. 19, n. 4, p. 246–256, abr. 2016. ISSN 2152-2715, 2152-2723. Available from Internet: <http://www.liebertpub.com/doi/10.1089/cyber.2014.0682>.

MOGLIA, A. et al. A Systematic Review of Virtual Reality Simulators for Robot-assisted Surgery. **European Urology**, v. 69, n. 6, p. 1065–1080, jun. 2016. ISSN 0302-2838. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S030228381500929X>.

MOHAMMADI, R. et al. Effects of Virtual Reality Compared to Conventional Therapy on Balance Poststroke: A Systematic Review and Meta-Analysis. **Journal of Stroke and Cerebrovascular Diseases**, v. 28, n. 7, p. 1787–1798, jul. 2019. ISSN 1052-3057. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1052305719301648>.

MORENO, A. et al. A systematic review of the use of virtual reality and its effects on cognition in individuals with neurocognitive disorders. **Alzheimer's & Dementia: Translational Research & Clinical Interventions**, v. 5, n. 1, p. 834–850, jan. 2019. ISSN 2352-8737. Publisher: John Wiley & Sons, Ltd. Available from Internet: <https://alz-journals.onlinelibrary.wiley.com/doi/full/10.1016/j.trci.2019.09.016>.

MUNZNER, T. **Visualization analysis and design**. Boca Raton: CRC Press, Taylor & Francis Group, CRC Press is an imprint of the Taylor & Francis Group, an informa business, 2015. (A.K. Peters visualization series). ISBN 978-1-4665-0891-0.

NEGUŢ, A. et al. Task difficulty of virtual reality-based assessment tools compared to classical paper-and-pencil or computerized measures: A meta-analytic approach. **Computers in Human Behavior**, v. 54, p. 414–424, jan. 2016. ISSN 0747-5632. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0747563215301059>.

NEUMANN, D. L. et al. A systematic review of the application of interactive virtual reality to sport. **Virtual Reality**, v. 22, n. 3, p. 183–198, sep. 2018. ISSN 1434-9957. Available from Internet: <https://doi.org/10.1007/s10055-017-0320-5>.

NG, Y.-L. et al. Effectiveness of virtual and augmented reality-enhanced exercise on physical activity, psychological outcomes, and physical performance: A systematic review and meta-analysis of randomized controlled trials. **Computers in Human Behavior**, v. 99, p. 278–291, oct. 2019. ISSN 0747-5632. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0747563219302018>.

NOBRE, C. et al. The State of the Art in Visualizing Multivariate Networks. **A. Lex**, p. 26, 2019.

NOCENTINI, A.; ZAMBUTO, V.; MENESINI, E. Anti-bullying programs and Information and Communication Technologies (ICTs): A systematic review. **Aggression and Violent Behavior**, v. 23, p. 52–60, jul. 2015. ISSN 1359-1789. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1359178915000749>.

NOUR, M. et al. A Narrative Review of Social Media and Game-Based Nutrition Interventions Targeted at Young Adults. **Journal of the Academy of Nutrition and Dietetics**, v. 117, n. 5, p. 735–752.e10, may 2017. ISSN 2212-2672. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S2212267216315490>.

NUSRAT, S.; KOBOUROV, S. The State of the Art in Cartograms. **Comput. Graph. Forum**, 2016.

PARSONS, T. D. Virtual Simulations and the Second Life Metaverse: Paradigm Shift in Neuropsychological Assessment. In: ZAGALO, N.; MORGADO, L.; BOA-VENTURA, A. (Ed.). **Virtual Worlds and Metaverse Platforms: New Communication and Identity Paradigms**. IGI Global, 2012, (Advances in Social Networking and

Online Communities). ISBN 978-1-60960-854-5 978-1-60960-855-2. Available from Internet: <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-854-5>.

PARSONS, T. D.; RIZZO, A. A. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. **Journal of Behavior Therapy and Experimental Psychiatry**, v. 39, n. 3, p. 250–261, sep. 2008. ISSN 00057916. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0005791607000456>.

PEREIRA, C. R. et al. A survey on computer-assisted Parkinson's Disease diagnosis. **Artificial Intelligence in Medicine**, v. 95, p. 48–63, abr. 2019. ISSN 0933-3657. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0933365717305663>.

PERLMAN, A.; SACKS, R.; BARAK, R. Hazard recognition and risk perception in construction. **Safety Science**, v. 64, p. 22–31, abr. 2014. ISSN 09257535. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0925753513002877>.

PERROCHON, A. et al. Exercise-based games interventions at home in individuals with a neurological disease: A systematic review and meta-analysis. **Annals of Physical and Rehabilitation Medicine**, v. 62, n. 5, p. 366–378, sep. 2019. ISSN 1877-0657. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1877065719300600>.

PFANDLER, M. et al. Virtual reality-based simulators for spine surgery: a systematic review. **The Spine Journal**, v. 17, n. 9, p. 1352–1363, sep. 2017. ISSN 1529-9430. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1529943017302085>.

PIMENTEL, K.; TEIXEIRA, K. Virtual reality: through the new looking glass. **Choice Reviews Online**, v. 30, n. 09, p. 30–5051–30–5051, may 1993. ISSN 0009-4978, 1523-8253. Available from Internet: <http://choicereviews.org/review/10.5860/CHOICE.30-5051>.

PIROLLI, P.; CARD, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: MCLEAN, VA, USA. **Proceedings of international conference on intelligence analysis**. [S.l.], 2005. v. 5, p. 2–4.

PRIBEANU, C.; BALOG, A.; IORDACHE, D. D. Measuring the perceived quality of an AR-based learning application: a multidimensional model. **Interactive Learning Environments**, v. 25, n. 4, p. 482–495, may 2017. ISSN 1049-4820, 1744-5191. Available from Internet: <https://www.tandfonline.com/doi/full/10.1080/10494820.2016.1143375>.

QUORA. **The Difference Between Virtual Reality, Augmented Reality And Mixed Reality**. 2018. Section: Tech. Available from Internet: <https://www.forbes.com/sites/quora/2018/02/02/the-difference-between-virtual-reality-augmented-reality-and-mixed-reality/>.

RAND, D. et al. Comparison of Two VR Platforms for Rehabilitation: Video Capture versus HMD. **Presence: Teleoperators and Virtual Environments**,

v. 14, n. 2, p. 147–160, abr. 2005. ISSN 1054-7460. Available from Internet: <https://direct.mit.edu/pvar/article/14/2/147-160/18536>.

RAVI, D. K.; KUMAR, N.; SINGHI, P. Effectiveness of virtual reality rehabilitation for children and adolescents with cerebral palsy: an updated evidence-based systematic review. **Physiotherapy**, v. 103, n. 3, p. 245–258, sep. 2017. ISSN 0031-9406. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0031940616300645>.

REPETTO, C.; RIVA, G. From virtual reality to interreality in the treatment of anxiety disorders. **Neuropsychiatry**, v. 1, n. 1, p. 31–43, feb. 2011. ISSN 1758-2008. Available from Internet: <http://www.futuremedicine.com/doi/10.2217/npy.11.5>.

RHEINGOLD, H. **Virtual reality**. New York: Summit Books, 1991. ISBN 978-0-671-69363-3.

RIOS, N.; NETO, M. G. d. M.; SPÍNOLA, R. O. A tertiary study on technical debt: Types, management strategies, research trends, and base information for practitioners. **Information and Software Technology**, v. 102, p. 117–145, oct. 2018. ISSN 09505849. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0950584918300946>.

RIVA, G. (Ed.). **Virtual reality in neuro-psycho-physiology: cognitive, clinical and methodological issues in assessment and rehabilitation**. Amsterdam ; Washington, D.C. : Tokyo: IOS Press ; Ohmsha, 1997. (Studies in health technology and informatics, v. 44). ISBN 978-90-5199-364-6 978-4-274-90207-9.

ROCHLEN, L. R.; LEVINE, R.; TAIT, A. R. First-Person Point-of-View–Augmented Reality for Central Line Insertion Training: A Usability and Feasibility Study. **Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare**, v. 12, n. 1, p. 57–62, feb. 2017. ISSN 1559-713X, 1559-2332. Available from Internet: <https://journals.lww.com/01266021-201702000-00009>.

ROONEY, M. K. et al. Simulation as More Than a Treatment-Planning Tool: A Systematic Review of the Literature on Radiation Oncology Simulation-Based Medical Education. **International Journal of Radiation Oncology*Biology*Physics**, v. 102, n. 2, p. 257–283, oct. 2018. ISSN 0360-3016. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0360301618309167>.

ROSE, F. D. et al. Learning and Memory in Virtual Environments: A Role in Neurorehabilitation? Questions (and Occasional Answers) from the University of East London. **Presence: Teleoperators and Virtual Environments**, v. 10, n. 4, p. 345–358, aug. 2001. ISSN 1054-7460. Available from Internet: <https://direct.mit.edu/pvar/article/10/4/345-358/18317>.

ROTHBAUM, B. O. et al. Virtual Reality Exposure Therapy and Standard (in Vivo) Exposure Therapy in the Treatment of Fear of Flying. **Behavior Therapy**, v. 37, n. 1, p. 80–90, mar. 2006. ISSN 00057894. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0005789406000104>.

RT ET AL., d. C. Virtual Reality Exposure in the Treatment of Fear of Flying. **Aviation, Space, and Environmental Medicine**, 2009. ISSN 00956562. Available from Internet: <http://www.ingentaconnect.com/content/10.3357/ASEM.2277.2008>.

SANTOS, M. E. C. et al. Augmented Reality Learning Experiences: Survey of Prototype Design and Evaluation. **IEEE Transactions on Learning Technologies**, v. 7, n. 1, p. 38–56, jan. 2014. ISSN 1939-1382. Available from Internet: <http://ieeexplore.ieee.org/document/6681863/>.

SAVRAN, M. M. et al. Training and Assessment of Hysteroscopic Skills: A Systematic Review. **Journal of Surgical Education**, v. 73, n. 5, p. 906–918, sep. 2016. ISSN 1931-7204. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1931720416300198>.

SCAPIN, S. et al. Virtual Reality in the treatment of burn patients: A systematic review. **Burns**, v. 44, n. 6, p. 1403–1416, sep. 2018. ISSN 0305-4179. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0305417917306022>.

SCHEFFLER, M. et al. Efficacy of non-pharmacological interventions for procedural pain relief in adults undergoing burn wound care: A systematic review and meta-analysis of randomized controlled trials. **Burns**, v. 44, n. 7, p. 1709–1720, nov. 2018. ISSN 0305-4179. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0305417917306599>.

SCHÖTTLER, S. et al. Visualizing and Interacting with Geospatial Networks: A Survey and Design Space. **Computer Graphics Forum**, v. 40, n. 6, p. 5–33, 2021. ISSN 1467-8659. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14198. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14198>.

SCHULTHEIS, M. T.; HIMELSTEIN, J.; RIZZO, A. A. Virtual Reality and Neuropsychology: Upgrading the Current Tools. **Journal of Head Trauma Rehabilitation**, v. 17, n. 5, p. 378–394, oct. 2002. ISSN 0885-9701. Available from Internet: <http://journals.lww.com/00001199-200210000-00002>.

SCHULTHEIS, M. T.; RIZZO, A. A. The application of virtual reality technology in rehabilitation. **Rehabilitation Psychology**, v. 46, n. 3, p. 296–311, 2001. ISSN 0090-5550. Available from Internet: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0090-5550.46.3.296>.

SCHULZ, H.-J. Treevis.net: A Tree Visualization Reference. **IEEE Computer Graphics and Applications**, v. 31, n. 6, p. 11–15, nov. 2011. ISSN 1558-1756. Conference Name: IEEE Computer Graphics and Applications.

SEE, K. W. M. et al. Evidence for Endovascular Simulation Training: A Systematic Review. **European Journal of Vascular and Endovascular Surgery**, v. 51, n. 3, p. 441–451, mar. 2016. ISSN 1078-5884. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1078588415007601>.

SHARP, H. **Interaction design 5e**. Indianapolis, IN: John Wiley and Sons, 2019. ISBN 978-1-119-54725-9.

SHARPLES, S. et al. Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems. **Displays**, v. 29, n. 2, p. 58–69, mar. 2008. ISSN 01419382. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S014193820700100X>.

SHERMAN, W. R.; CRAIG, A. B. **Understanding Virtual Reality: Interface, Application, and Design**. 1ª edição. ed. Amsterdam ; Boston: Morgan Kaufmann, 2002. ISBN 978-1-55860-353-0.

SLATER, M.; WILBUR, S. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. **Presence: Teleoperators and Virtual Environments**, v. 6, n. 6, p. 603–616, dec. 1997. ISSN 1054-7460. Available from Internet: <https://direct.mit.edu/pvar/article/6/6/603-616/18157>.

SUBRAMANIAN, S. K.; PRASANNA, S. S. Virtual Reality and Noninvasive Brain Stimulation in Stroke: How Effective Is Their Combination for Upper Limb Motor Improvement?—A Meta-Analysis. **PM&R**, v. 10, n. 11, p. 1261–1270, 2018. ISSN 1934-1563. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.pmrj.2018.10.001. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1016/j.pmrj.2018.10.001>.

SUH, A.; PROPHET, J. The state of immersive technology research: A literature analysis. **Computers in Human Behavior**, v. 86, p. 77–90, sep. 2018. ISSN 0747-5632. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0747563218301857>.

TAN, B.-L.; LEE, S.-A.; LEE, J. Social cognitive interventions for people with schizophrenia: A systematic review. **Asian Journal of Psychiatry**, v. 35, p. 115–131, jun. 2018. ISSN 1876-2018. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1876201815300496>.

TAY, C.; KHAJURIA, A.; GUPTE, C. Simulation training: A systematic review of simulation in arthroscopy and proposal of a new competency-based training framework. **International Journal of Surgery**, v. 12, n. 6, p. 626–633, jun. 2014. ISSN 1743-9191. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1743919114000909>.

THOMSEN, A. S. S. et al. Update on Simulation-Based Surgical Training and Assessment in Ophthalmology: A Systematic Review. **Ophthalmology**, v. 122, n. 6, p. 1111–1130.e1, jun. 2015. ISSN 0161-6420. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0161642015001682>.

TROCHIM, W. M. K.; DONNELLY, J. P. **Research methods knowledge base**. 3. ed. ed. Mason, Ohio: Cengage Learning, 2008. OCLC: 845217982. ISBN 978-1-59260-291-9 978-1-59260-290-2.

TURNER, W. A.; CASEY, L. M. Outcomes associated with virtual reality in psychological interventions: where are we now? **Clinical Psychology Review**, v. 34, n. 8, p. 634–644, dec. 2014. ISSN 0272-7358. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0272735814001391>.

VIÑAS-DIZ, S.; SOBRIDO-PRIETO, M. Virtual reality for therapeutic purposes in stroke: A systematic review. **Neurología (English Edition)**, v. 31, n. 4, p. 255–277, may 2016. ISSN 2173-5808. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S2173580816300062>.

VOGT, S. et al. Virtual reality interventions for balance prevention and rehabilitation after musculoskeletal lower limb impairments in young up to middle-aged adults: A comprehensive review on used technology, balance outcome measures and observed effects. **International Journal of Medical Informatics**, v. 126, p. 46–58, jun. 2019. ISSN 1386-5056. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1386505618303010>.

WEISS, P. L. et al. Virtual reality in neurorehabilitation. In: SELZER, M. et al. (Ed.). **Textbook of Neural Repair and Rehabilitation**. Cambridge: Cambridge University Press, 2006. p. 182–197. ISBN 978-0-511-54507-8. Available from Internet: <https://www.cambridge.org/core/product/identifier/CBO9780511545078A028/type/book_part>.

WEISS, P. L. T.; TIROSH, E.; FEHLINGS, D. Role of Virtual Reality for Cerebral Palsy Management. **Journal of Child Neurology**, v. 29, n. 8, p. 1119–1124, aug. 2014. ISSN 0883-0738, 1708-8283. Available from Internet: <http://journals.sagepub.com/doi/10.1177/0883073814533007>.

WINN, W. **A Conceptual Basis for Educational Applications**. 1993. Available from Internet: <http://www.hitl.washington.edu/research/learning_center/winn/winn-paper.html~>.

WOJCIECHOWSKI, R.; CELLARY, W. Evaluation of learners' attitude toward learning in ARIES augmented reality environments. **Computers & Education**, v. 68, p. 570–585, oct. 2013. ISSN 03601315. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0360131513000535>.

ZENG, W.; RICHARDSON, A. Adding Dimension to Content: Immersive Virtual Reality for e-Commerce. p. 8, 2016.

APPENDIX A — O PANORAMA DE AVALIAÇÃO EM XR: REVISÃO TERCIÁRIA E VISUALIZAÇÕES

Palavras-chave: avaliação, realidade virtual, realidade aumentada, realidade mista, revisão sistemática, revisão terciária

Aplicações de realidade estendida (XR) - termo que abrange Realidade Virtual, Realidade Aumentada e Realidade Mista - estão encontrando seu caminho em vários domínios em um ritmo acelerado. Cada domínio de aplicação tem diferentes motivações para empregar XR e diferentes critérios segundo os quais avaliar o sucesso do uso de XR. Para entender o uso de XR em áreas diversas, várias revisões da literatura - de diferentes áreas - descrevem aplicações de XR e como elas são avaliadas. No entanto, nem sempre há uma definição clara do que é XR para cada área específica e, quando há, esta pode diferir substancialmente daquela usada em outras áreas. Essa falta de consenso sobre uma definição torna difícil comparar os esforços de pesquisa em XR entre as áreas e aprender com eles. Por meio de uma revisão sistemática terciária da literatura, foram analisados 81 destes artigos de revisão publicados em diversos domínios de aplicação para construir um resumo abrangente do estado atual da pesquisa em XR que envolve avaliação de aplicações de XR. A pesquisa foi baseada no entendimento de (i) como XR é definida? (ii) por que XR é empregada? (iii) como XR é avaliada? (iv) quais as principais críticas e caminhos apontados para pesquisas futuras delineados pelos artigos estudados? (v) quão boas são as revisões, segundo os critérios da *Database of Abstracts of Reviews of Effects* (DARE)? Essas perguntas guiaram a coleta de dados a partir da leitura de cada uma das revisões.

As revisões foram categorizadas em dez diferentes domínios: Simuladores - sistemas VR usados em Medicina para o treinamento de cirurgiões; Aprendizado - que compreende o uso de XR em todos os níveis de educação; Psicologia - aplicações VR usadas para tratamentos psicológicos, especialmente terapia de exposição; Rehabilitação pós-derrame - aplicações VR que visam diminuir sintomas físicos em pacientes de derrame; Cognição - estudam o uso de VR como tratamento ou forma de diagnóstico para doenças neurocognitivas; Cirurgia - focam no uso de XR como um auxílio na sala de cirurgia; Alívio da dor - foca no uso de VR no tratamento de dor durante procedimentos médicos; Prevenção física - estuda o uso de XR para estimular o exercício em adultos saudáveis; Múltiplas áreas - revisões que estudam XR em mais de um domínio; e Indústria - foca no uso de AR para processos industriais.

Os resultados foram apresentados por meio de visualizações dos dados coletados - representando o panorama de avaliação em XR - e foi descrito o estado da pesquisa em XR em cada um dos domínios encontrados. Para elaborar as visualizações, analisou-se o estado da arte da visualização de dados de revisões. A partir dessa análise, foi proposta uma nova forma de visualização, dada a estrutura dos dados coletados - que não poderia ser visualizada apropriadamente pelas técnicas já existentes.

Os pontos fortes encontrados na pesquisa atual em XR são: o interesse em XR está aumentando e se espalhando por diversas áreas, o que é explicitado pelo número crescente de revisões ano a ano; em alguns domínios, como Psicologia e Aprendizado, já existem fortes evidências da eficácia de XR. Já as lacunas identificadas são: definições inconsistentes ou ausentes de XR; limitações da pesquisa atual, que não costuma comparar diferentes tipos de XR entre si, ou fatores específicos de XR (por exemplo, imersão) em diferentes níveis, para estabelecer mais claramente uma relação causal entre tal fator e os resultados encontrados; poucos dos efeitos e medidas utilizados para avaliar XR nos diferentes domínios são relacionados à usabilidade, IHC ou ergonomia - inversamente, a maioria (91%) dos pares efeito-medida aparecem em apenas uma revisão, ou seja, são bastante específicos de cada domínio; finalmente, muitas das revisões pedem que os estudos primários tenham designs experimentais mais fortes - por exemplo, aumentando o número de participantes.

Foram identificados três caminhos para pesquisa futura em XR. Primeiro, avaliar humanos usando XR - aproveitando-se da abundância de sensores que sistemas de XR já possuem (por exemplo câmeras e sensores inerciais em um HMD) - XR poderia ser usado para avaliar ações humanas de forma não intrusiva e objetiva, como a performance de um cirurgião, o progresso na reabilitação de um paciente de derrame ou a ergonomia de uma tarefa realizada por um trabalhador em uma fábrica.

Segundo, usar XR como um ambiente controlado para estudar o comportamento humano - XR pode ser utilizado para simular situações que seriam perigosas ou antiéticas se performadas na realidade, como analisar como um médico conduz uma cirurgia privado de sono sob o efeito de drogas, para estudar seu impacto na sua performance.

Finalmente, conclue-se que a pesquisa em XR deve sair da pesquisa sobre eficácia em direção à pesquisa sobre eficiência. Para isso, as áreas devem passar por quatro etapas distintas: definir o que é XR, descobrir se XR é efetivo, descobrir se XR é eficiente, e finalmente descobrir *como* XR funciona.