Multimodal Technologies and Interaction

MDPI

Article

User-Centered Evaluation Framework to Support the Interaction Design for Augmented Reality Applications

Andrea Picardi and Giandomenico Caruso *

Department of Mechanical Engineering, Politecnico di Milano, Via La Masa 1, 20156 Milano, Italy; andrea.picardi@polimi.it

* Correspondence: giandomenico.caruso@polimi.it; Tel.: +39-02-2399-8094

Abstract: The advancement of Augmented Reality (AR) technology has been remarkable, enabling the augmentation of user perception with timely information. This progress holds great promise in the field of interaction design. However, the mere advancement of technology is not enough to ensure widespread adoption. The user dimension has been somewhat overlooked in AR research due to a lack of attention to user motivations, needs, usability, and perceived value. The critical aspects of AR technology tend to be overshadowed by the technology itself. To ensure appropriate future assessments, it is necessary to thoroughly examine and categorize all the methods used for AR technology validation. By identifying and classifying these evaluation methods, researchers and practitioners will be better equipped to develop and validate new AR techniques and applications. Therefore, comprehensive and systematic evaluations are critical to the advancement and sustainability of AR technology. This paper presents a theoretical framework derived from a cluster analysis of the most efficient evaluation methods for AR extracted from 399 papers. Evaluation methods were clustered according to the application domains and the human–computer interaction aspects to be investigated. This framework should facilitate rapid development cycles prioritizing user requirements, ultimately leading to groundbreaking interaction methods accessible to a broader audience beyond research and development centers.

Keywords: augmented reality; human–computer interaction; user evaluation; methods and tools; evaluation framework; user testing

Citation: Picardi, A.; Caruso, G. User-Centered Evaluation Framework to Support the Interaction Design for Augmented Reality Applications. *Multimodal Technol. Interact.* **2024**, *8*, 41. https://doi.org/10.3390/mti8050041

Academic Editor: Cristina Portales

Received: 14 April 2024
Revised: 5 May 2024
Accepted: 10 May 2024
Published: 14 May 2024

1. Introduction

Augmented Reality (AR) has been a promising technology for interaction design for three decades, with the potential to change our actions, perceptions, and world experiences. With a long history of developments, the last years have seen a steady growth of advancements [1], approaching what is commonly referred to as the "next-generation interface" [2,3]. Barriers between the real world and digital information are broken down, and the surrounding environment becomes the new medium of interaction by representing a significant leap in human–computer interaction after the graphical user interface [4].

While researchers have made significant strides in enabling AR to overcome technological limitations and adopt new interaction paradigms, several technical and user-related challenges remain. Notably, most prior research has primarily focused on hardware and software challenges, with only limited attention given to the user-centric aspects within this domain. This underscores the need for a shift in focus towards user-centric research in AR, as user experience is a crucial factor in effective design and evaluation.

The research process in AR heavily relies on technology, resulting in new technologies and prototypes at the beginning that may not fully address user problems. This process often prioritizes technological advancement over user needs, sometimes resulting in useless implementations. Most emerging technologies prioritize technical issues over user involvement, but user experience is crucial for effective design and evaluation [5]. Similarly,

in [6], the authors found a limited number of publications that discuss human–computer interaction (HCI), and even those that did conducted user-based experiments that were minimal. They concluded that the lack of formal user evaluations is due to a lack of knowledge among researchers on which tools to use for their specific situations.

Over the past decade, there has been a moderate improvement in user-centered research, design influenced by user input, and previous evaluations conducted with actual users [7,8]. However, there remains a requirement for additional research that prioritizes the human-centered approach in this domain, aimed at creating dependable evaluation methods and guiding principles. Efforts are required to rearrange and re-evaluate the existing state of the art in developing AR systems. This will provide a solid foundation for future user evaluations while considering end-users importance, needs, emotions, and desires. These aspects are often overlooked in the technology-driven development of this field. Still, they play a fundamental role in determining the effectiveness of an AR product and, consequently, its usage [9].

A helpful approach to improve user-centered research for AR technology is to classify and update previous assessments in a practical format. This will offer guidance and understanding for future researchers to enhance their evaluations of AR systems. Ultimately, this will lead to a more widespread user adoption and the use of this innovative technology. It is crucial to review the achievements within each domain of application retrospectively. Furthermore, we should consider researchers' tools and methods in their respective fields. By understanding the historical practices and approaches used to address this problem, we can enhance the evaluation techniques employed within AR. This exploration can provide valuable insights into past efforts and successful practices, ultimately defining the current state of the art.

However, being aware of prior research is insufficient for making informed evaluations in the future. Understanding the specific aspects of HCI investigated using these tools is equally crucial. This understanding is pivotal in deciphering the rationale behind their utilization and identifying the focal points of AR research that garnered the most attention in the past. Such insights will aid in the determination of whether forthcoming evaluations should continue to emphasize the same facets or allocate resources to different areas. Although there is no universal agreement on the tools used to investigate AR interaction across various fields, more and more scholars are highlighting the importance of prioritizing technology development and user interaction evaluation. This approach is essential for advancing AR technologies and creating a more sophisticated future [6,9].

Following a comprehensive examination of these aspects, the research presented in this paper involves the organization of the acquired data and the formulation of an efficient framework to achieve these three primary objectives:

1. Offering a comprehensive exploration of the assessment techniques of different HCI issues applied in AR;
2. Assisting researchers in selecting appropriate methods based on the characteristics of technologies and application domains;
3. Providing insight into the essential role of user evaluations in ensuring the success and widespread acceptance of forthcoming AR technologies and applications.

After the introduction, Section 2 of the paper discusses the importance of applying an effective evaluation strategy for AR applications and the current limitations. Section 3 provides all the elements elaborated to develop the proposed evaluation framework. Section 4 reports the analysis results of the papers identified as the most relevant for this study. Section 5 presents the framework implementation and its practical usage. Finally, Section 6 discusses the achieved results, while Section 7 concludes the work.

2. Background

In the design process, user evaluation is crucial to gather and analyze data on how users interact with a system or artifact. This helps identify potential problems during development or in future iterations [10]. It is important to note that user evaluation is

not the same as usability evaluation, although the terms are often used interchangeably. Usability testing involves selecting representative tasks based on learnability, efficiency, memorability, errors, and satisfaction [11]. However, there are also other models in use. These attributes are assessed and appraised through the active involvement of actual users, with the primary aim of gathering essential information to address the issues uncovered during the testing process. However, their primary focus remains on the system's usability, and the definition of "usability" may vary depending on the model in use. There are various situations where user evaluations can serve different purposes. For example, they can be used for comparing a new interaction technique in terms of user efficiency or accuracy, examining user behavior and interaction with a new prototype, or investigating how a new system facilitates collaboration, and these are just a few of the many possibilities. Various evaluation methods are employed throughout the development cycle, starting from the initial stages, like the ideation phase, where approaches such as Heuristic evaluation [12] and Wizard of Oz [13] techniques are utilized, to the concluding stages, where tests involving physical prototypes and actual users are carried out. It is crucial to explicitly define the objectives of each evaluation study, as conducting a trial with the prototype or employing evaluation techniques designed for usability testing does not automatically indicate the execution of a usability test [14].

Assessments can start at the beginning of the development process and encompass a broad spectrum of processes, from the initial phases of prototype creation to fine-tuning and enhancing a nearly completed design. Conducting periodic and informal tests to implement iterative enhancements that quickly address minor usability and design issues can be highly beneficial throughout the project.

It is important to note that validating a prototype does not necessarily provide insights for its everyday usage. During a user evaluation, various factors should be considered, such as the social context, overall usefulness of the system, and ease of use. In [15], the authors have raised concerns about the potential risk of using evaluation methods in the wrong context. While the process remains always valid, if it is used appropriately, its application in the wrong context makes the results useless and sometimes misleading. Consequently, the appropriate evaluation methods should be based on the specific application problems or well-defined research questions. With this aim, the framework proposed in this paper should assist researchers in selecting an evaluation method appropriate to their study.

3. Framework Elements

The framework has been elaborated according to the elements characterizing the AR solution under investigation. These elements are Application Domains, Investigated HCI aspects, and evaluation methods and tools. This paragraph describes how these elements have been analyzed and clustered to define the specific taxonomies of the proposed evaluation framework.

3.1. Application Domains

The proposed framework's starting point is identifying the application domain of the AR system. According to the literature, we adopted (and expanded) the taxonomy provided in [7,14]. These classifications have been reorganized and extended to better describe the application's different domains according to this study's aims. Besides the domains, we further extend the domain definition by including subdomains to describe each study's domain in more detail. The list of these domains and subdomains with their definition is shown in Table 1.

Table 1. Defined domains, subdomains, and descriptions to elaborate the framework. * New domains not included in the previous taxonomies [7,14].

Domains	Subdomains	Descriptions
Business and Services	Advertising/Product preview, Fashion/Makeup, Retail	AR applications to increase consumer awareness and brand recognition.
Communication and Telepresence	Remote Help, Telepresence and Remote collaboration, Telepresence Surgery	AR applications foster collaboration by enhancing the remote sense of presence.
Cultural and Tourism	Commercial exploration and discovery, Heritage exploration and discovery, Museum and Exhibitions	AR systems to enhance museum exhibitions, heritage explorations, and tourism.
Education and Training	Design, Engineering and Architecture, History, Languages, Music, Orientation, Physical Activities, Science subjects, Serious games, Special needs education	AR solutions stimulate the learning process by making teaching more interactive and engaging.
Entertainment	Gaming, Music, Narrative experience	AR applications that include explicit ludic components like games, narratives, experiences, toys, etc.
Field Operations *	Archaeological, Crime Scene Investigation, Military Operations, On-site planning/maintenance	A new domain was added due to the many studies conducted in the field that focused on some operational work.
Generic Interface *	Calibration, Collaboration, Info presentation/visualization, Interactions and Ergonomics, Perception, Tangible Interface	AR interfaces without a specific application field were thus evaluated from a generic perspective.
Health Care and Medicine	Elderly, Disables Help, Emergency, Personal Help, Phobia Treatment, Rehabilitation, Surgery, Training	AR solutions to help both patients and healthcare workers, including therapeutic, rehabilitation, and assistance fields.
Industry	Assembly, Design and Engineering, Logistics, Maintenance, Manufacturing, Training	AR technologies to design and validate prototypes in the early phases of work for maintenance, manufacturing, and logistical support for goods, buildings, and services.
Navigation and Driving	Driving, Info/Annotations AR and Remote viewing, Inside orientation and space navigation, Outside orientation and space navigation, Remote orientation and navigation	AR systems support users in driving or piloting vehicles, navigating the environment, and informing them about their surroundings.
Other	Expectations/acceptance, Immersion/motivation, Privacy, State of the art, Human/Robot/AI Interaction Security	All the entries that could not be classified in the other domains

3.2. Investigated HCI Aspects

Within the domain of HCI, evaluations are carried out for various purposes, such as evaluating the functionality of an interface, comprehending the intention behind a user's actions, assessing user perception, and observing their conduct. For more clarity, we have homed in on specific aspects that have already undergone extensive scrutiny in the existing literature.

These three HCI aspects were identified in [6]: Human Perception and Cognition in AR (effects of AR display viewing conditions, display hardware specifications, and depth perception on alternative rendering techniques), Performance (how users perform tasks when using AR applications within specific application domains), and Interaction and Communications between users (works centered on collaborating users, and how they share the same AR space at the same time).

Subsequently, in [16], the authors introduced system usability/system design evaluation, which does not necessarily involve assessing user task performance but instead focuses on identifying usability issues with the device. Usability encompasses ease of use, usefulness, learnability, satisfaction, and comfort.

In [17], User Experience (UX) substituted usability with the definition of "subjective user issues, such as technology preference, effect, perceptual and physical experiences", following the definitions introduced in [18,19]. However, it is less common to investigate UX using controlled experimental methods, which conduct them to separate this category into two sub-groups: formal UX evaluations (involving controlled experiments with a fixed sample of users and collected participants' experiences with tools like structured surveys/questionnaires) and informal UX evaluations (involving unstructured interviews or observations with a casual sample of potential users or domain experts) [20].

They discovered that many experiments in other categories included methods that could be classified as UX evaluations, such as a questionnaire at the end of a task performance assessment or an unstructured interview after a collaborative section. The definition of UX is broad and encompasses many aspects of the human experience, so using a single method to assess a single factor contributing to successful HCI is impossible. Therefore, these categorizations are relatively flexible and can accommodate evaluations that refer to more than one category. This finding was crucial in the parameters definition phase as it led to the realization that one evaluation method could be used to determine multiple HCI aspects and cover more evaluation goals than one.

After carefully examining the considerations and categorization mentioned above and reviewing the evaluation methods presented in previous research, we found the categorization they defended overly restrictive. This is due to the broad definition that these parameters identify and describe, particularly regarding UX. We were compelled to enlarge and reorganize the suggested categories according to these papers [21–23]. In these studies, the authors aimed to identify the aspects that could define a successful and enjoyable user interaction, as described in [24,25] for UX in the field of HCI. Following these studies, we adopted the following four macro categories of UX: Emotion, Meaning, Usability, and Usefulness.

This was necessary due to the complexity of UX, which lacks a standard definition [26]. Usability and UX are often considered separate concepts, but usability is crucial to the UX. Sometimes, UX is regarded as a subset of usability, as it focuses on the satisfaction aspect of usability [21]. This is also true for usefulness. Some see it as a descriptive component of usability, while others see it as a separate subject. [27].

We adopted the model identified in [25] to accommodate the various perspectives. This model effectively described each category and provided a comprehensive collection of examples that helped us create clear definitions for each category. Based on these definitions and the previous surveys, we identified 14 HCI aspects collected in Table 2 with their descriptions.

Table 2. Investigated HCI aspects.

Investigated HCI Aspects	Description
Collaboration and Communication	User interaction and communication between collaborating users without a specific application domain.
Education specific	Effectiveness of learning using AR.
Ergonomics, Loads, and Comfort	Mental and physical load, comfort, sickness, frustration, anxiety, or stress.
Interaction	Users' behaviors, interaction patterns and strategies, attentions, and actions during the tests.
Mixed background questions	Evaluation of generic aspects like technological comfort, general use, and subjective interest in the topic.

Table 2. *Cont.*

Investigated HCI Aspects	Description
Perception and Cognition	Perceptual effects of alternative rendering techniques, depth perception, etc.
Prototype focus	Evaluation of the AR system prototype: missing features, general opinions, feedback, and suggestions.
Task performance	Users' performance in their interactions with the system, e.g., time, errors, length traveled, etc.
Treatment Specific	Medical treatment, rehabilitation, or phobia study. It is tied with the Health Care and Medicine domain.
UX—Emotion	All the affective components the user experiences, e.g., satisfaction, joy of usage, pleasure, excitement, amusement, etc.
UX—Meaning	The users expected long-term personal relationships with the system or product addressed, the society, personal beliefs, self-expression, and more.
UX—Usability	How well users can use a product according to different aspects, e.g., ease of use, performance, efficiency, error avoidance, learnability, memorability, etc.
UX—Usefulness	Ability perceived by the user of the system or product addressed to accomplish the user's pre-determinate goals.
Other	Elements that are not categorized into other groups.

3.3. Evaluation Methods and Tools

To cluster and organize the users' evaluation tools and methods, we closely followed the taxonomy defined in [27]. This taxonomy, designed to encompass the diverse range of techniques employed in evaluating AR systems, exhibits commendable flexibility. Its adaptability renders it suitable for widespread adoption. However, a critique arises regarding the rigidity observed in categorizing evaluation methods within the associated documents. This limitation prompted an extension of the taxonomy, introducing supplementary subtypes. These subtypes refine the taxonomy, facilitating a more precise depiction of evaluation procedures by incorporating the diverse tools utilized, e.g., questionnaire typology (NASA-TLX, SUS, home-made, etc.), rating scale (Likert, EZ scale, FSS, etc.), action logging type (system logs, tracking logs, etc.), etc. These subgroups are detailed for each domain within the implemented framework and are available in the Supplementary Materials of this paper. Table 3 describes all the methods identified as the most representative for evaluating AR applications.

Table 3. Classification and description of the evaluation methods.

Methods	Description
Conventional Test (Written/Oral)	Tests are carried out in an educational setting (e.g., schools or universities), where the teacher performs a test to understand the efficiency of the method taught (this is specific to the Education and Training domain).
Experts Review/Evaluation	One or more experts in the field evaluate the product or system addressed using cognitive or pluralistic walkthroughs, heuristic evaluations, product insights, etc.
Focus Group	Evaluations are conducted in a focus group environment, where users, stakeholders, or experts meet and discuss the prototype using different techniques.
Interviews	Evaluation is conducted orally, where an exterminator asks the users to answer questions about the product or system. This could be structured or unstructured, but the users would express their opinions qualitatively.
Observation	Evaluations conducted by the examiner, who indirectly observes the users interacting with or using the prototype, are usually qualitative.

Table 3. *Cont.*

Methods	Description
Question-answer-Protocol	A procedure where the examiner asks a question to the user, and the user should answer with a predetermined answer, such as Yes or No. These questions are asked while the user is interacting with the prototype.
Questionnaire	Surveys in which users must express their opinions following predetermined questions and answers. These questionnaires could be qualitative or quantitative.
Self-Reported/Diaries	Evaluations are where the users are asked to report about the use of the product or system during an extended period, so other assessments are not possible.
Think aloud/Shadowing	Evaluation involves following the users by an exterminator/facilitator who observes and reports their interactions and behaviors with the prototype. This differs from observation because the users know about the evaluator's presence.
User Action Logging	The system logs the user's actions and behaviors in this evaluation and later elaborates.
User Measurements	Evaluation where the user's physical attributes (such as heart rate, eye patterns, and so on. . .) are measured and tracked during the exam.

4. Paper Analysis Results

The framework has been elaborated by analyzing the 433 papers from three systematic reviews covering 2001 to 2018 [7,17,28]. These collections have been chosen based on three main reasons. Their primary objective was to examine the literature about evaluation techniques used in the field, which meant that the materials collected contained some form of evaluation. Secondly, they already included a categorization and sampling of evaluations closely aligned with our definitions. Lastly, they openly disclosed the papers that were analyzed, making the selection process of materials more accessible. After analyzing each paper's title, abstract, and contents, we found that 34 papers were duplicates, leaving 399 unique papers. Of those, 36 (8.3%) contained multiple evaluations, leading to 473 unique assessments. Of these evaluations, 29 are pre-studies conducted on a small sample of users. These papers underwent a thorough process to extract the framework base elements. Initially, the studies were read and then manually categorized based on the predetermined characteristics of our framework. Following this, a spreadsheet was generated to delineate the framework categories as column headings, while individual papers were itemized as rows. Subsequently, pertinent data were systematically inputted into each row. Employing the table pivot tool, the categorized data underwent efficient rearrangement, visualization, and analysis. After this phase, the principal constituents of the framework were crafted utilizing graphical and tabular depictions derived from the collected data.

As shown in Table 4, the number of application domains is not equally distributed among the analyzed papers. The top three domains with the highest occurrences represent over 60% of all the studies conducted.

Table 4. Distribution of the application domains within the analyzed papers.

Domains	Occurrences	Percentage
Generic Interface	164	34.7%
Education and Training	70	14.8%
Health Care and Medicine	60	12.7%
Industry	39	8.2%
Navigation and Driving	39	8.2%
Cultural and Tourism	26	5.5%
Entertainment	22	4.7%
Communication and Telepresence	16	3.4%
Field Operations	16	3.4%
Business and Services	11	2.3%
Other	10	2.1%
TOTAL	473	100%

The most representative generic interface domains are perception (15.4%), interactions and ergonomics (9.5%), info presentation/visualization (4.9%), and tangible interfaces (3%). This is because researchers primarily explore the general principles of perception and interaction in the initial decade of research rather than developing specific prototypes for particular application domains.

In the Education and Training domain, most of the studies relate to the subdomains Science subjects (3.8%), Design, Engineering, and Architecture (3.4%), and serious games (2.5%), which confirms the interest in exploiting AR as a medium to foster the learning of even complex content.

Health Care and Medicine focuses on training medical staff in delicate and dangerous situations (4%) or supporting surgeons (2.1%) by offering novel and minimally invasive visualization and interaction methods for surgical instruments and patients. Rehabilitation (3%) and phobia treatment (1.5%) also depended on the potential realism that could be enhanced through AR interfaces.

Industry and Navigation and Driving share a similar number of studies, mainly concentrated in maintenance (2.1%) and Design and Engineering for the first, while info/annotations AR and remote viewing (2.3%) and driving (2.1%) for the latter.

The Cultural and Tourism domain mainly evaluates AR applications for museum exhibitions (2.3%), while Entertainment primarily includes evaluations in the gaming subdomain (3.4%).

Finally, Communication and Telepresence, Field Operations, and Business and Services include limited studies in their corresponding subdomains. This could be due to the limited number of use cases that do not require continuous and extensive evaluations.

The investigated HCI aspects have a less polarized distribution, even though the first aspect (UX—Usability) is almost 15 times higher than the last one (Treatment Specific). Table 5 shows the overall instances and their relative percentages, representing their relevance in the analyzed papers. It is worth noting that the overall HCI occurrences (1566) are much higher than the studies (473) because each study could evaluate more than one HCI aspect.

Table 5. Distribution of the investigated HCI aspects considered for the framework and their relevance in occurrences and percentage.

Investigated HCI Aspects	Occurrences	Percentage
UX—Usability	288	18.39%
Perception/Cognition	213	13.60%
Prototype focus	190	12.13%
Task performance	181	11.56%
UX—Emotion	139	8.88%
Interaction	138	8.81%
Ergonomics/Load/Comfort	120	7.66%
UX—Usefulness	93	5.94%
Collaboration/Communication	51	3.26%
Mixed background questions	50	3.19%
Education specific	48	3.07%
UX—Meaning	31	1.98%
Treatment Specific	20	1.28%
other	4	0.26%
TOTAL	1566	100%

The analysis shows that usability was the most researched aspect (18.39%), followed by perception/cognition (13.6%) and prototype focus (12.13%). The first category is a significant area of interest for AR interfaces, and the latter is a common usability practice used to evaluate a system's efficiency. Upon examining the ergonomics, loads, comfort group, and prototype focus, we discovered a cluster of UX emotions (8.88%) focusing on amusement, motivation, and general interest in the system and technology. The interaction group is just

slightly behind (8.81%) in the UX-Usefulness (5.94%) category, which is essential for the success and adoption of new devices but is rarely investigated and considered in this type of assessment. We believe that this lack of effort spent on this category could be due to the technological push that drives the development of AR technologies and the resulting lack of user-centered approaches in AR systems [16]. The remaining aspects include Communication (3.26%), Mixed background questions (3.19%), and treatment-specific (1.28%). The UX-Meaning (1.98%) aspect is not explored frequently, possibly because of the newness of AR technology and the absence of consistent prior experiences.

Table 6 lists the methods used to investigate the different aspects of HCI and their relevance in the analyzed papers. Their overall occurrences (963) are still higher than the overall studies (473) but less than the methods (1566) because even if one study can include more than one method, a single method can be used to investigate more than one HCI aspect.

Table 6. List the methods considered in the framework and their relevance in terms of occurrence and percentage.

Methods	Occurrences	Percentage
Questionnaire	347	36.03%
User Action Logging	252	26.17%
Observations	119	12.36%
Interviews	94	9.76%
Question-answer-Protocol	52	5.40%
Conventional Test (Written/Oral)	29	3.01%
User Measurements	27	2.80%
Experts Review/Evaluation	20	2.08%
Focus Group	8	0.83%
Think aloud/Shadowing	8	0.83%
Self-Reported/Diaries	7	0.73%
TOTAL	963	100%

Despite the HCI aspects, the first four more frequent methods represent almost 85% of the occurrences. Questionnaires (36.03%) are the most used tool, followed by user action logging (26.17%), observations (12.36%), and interviews (9.36%). Few papers used other evaluation methods such as question-answer protocol (5.40%), conventional tests (3.01%), direct user measurements (2.8%), and expert reviews (2.08%). The remaining methods (focus groups, shadowing, think-aloud, diary studies co-discover, and cognitive walkthroughs) represent less than 3% of the analyzed papers.

These results align with those found in [27,29,30], where the authors attributed this unbalance of evaluation methods to the difficulties in their adoption due to a nonstandard and unclear definition. It is commonly believed that user action logging (including time and error recording) and direct observation (or analysis of recorded material) are the most frequently used methods in standard usability evaluation. These methods involve tracking user actions to measure task performance and using standardized questionnaires to assess usability.

The distribution of subtypes is noteworthy as there is an uneven distribution within the group of questionnaire inquiries, and the notable tools used were not standardized but customized for the situation. This is a common practice in prototype evaluation. However, the significant gap between the top two entries (custom-made and NASA TLX) indicates the lack of maturity of the evaluation methods for AR systems. The absence of standardized evaluation protocols and tools, combined with the many possibilities, makes it quite challenging to find standard evaluation tools that could be used in other fields and solutions.

Finally, the subtype methods used per single study varied; 49.1% of studies used only one subtype method, and 28.7% used two subtype methods. Less frequent were the cases with three (13.5%), four (5.3%), five (2.1%), and six (0.8%), and only five studies used seven methods (0.5%).

5. Framework Implementation and Usage

Based on the findings presented in Section 4, the framework addresses two main requirements. Firstly, it aims to showcase the primary user evaluation methods and HCI aspects previously utilized and proven effective in similar contexts. Secondly, it intends to assist in selecting the appropriate tool for new evaluations in the field of AR, thereby facilitating the entire evaluation process.

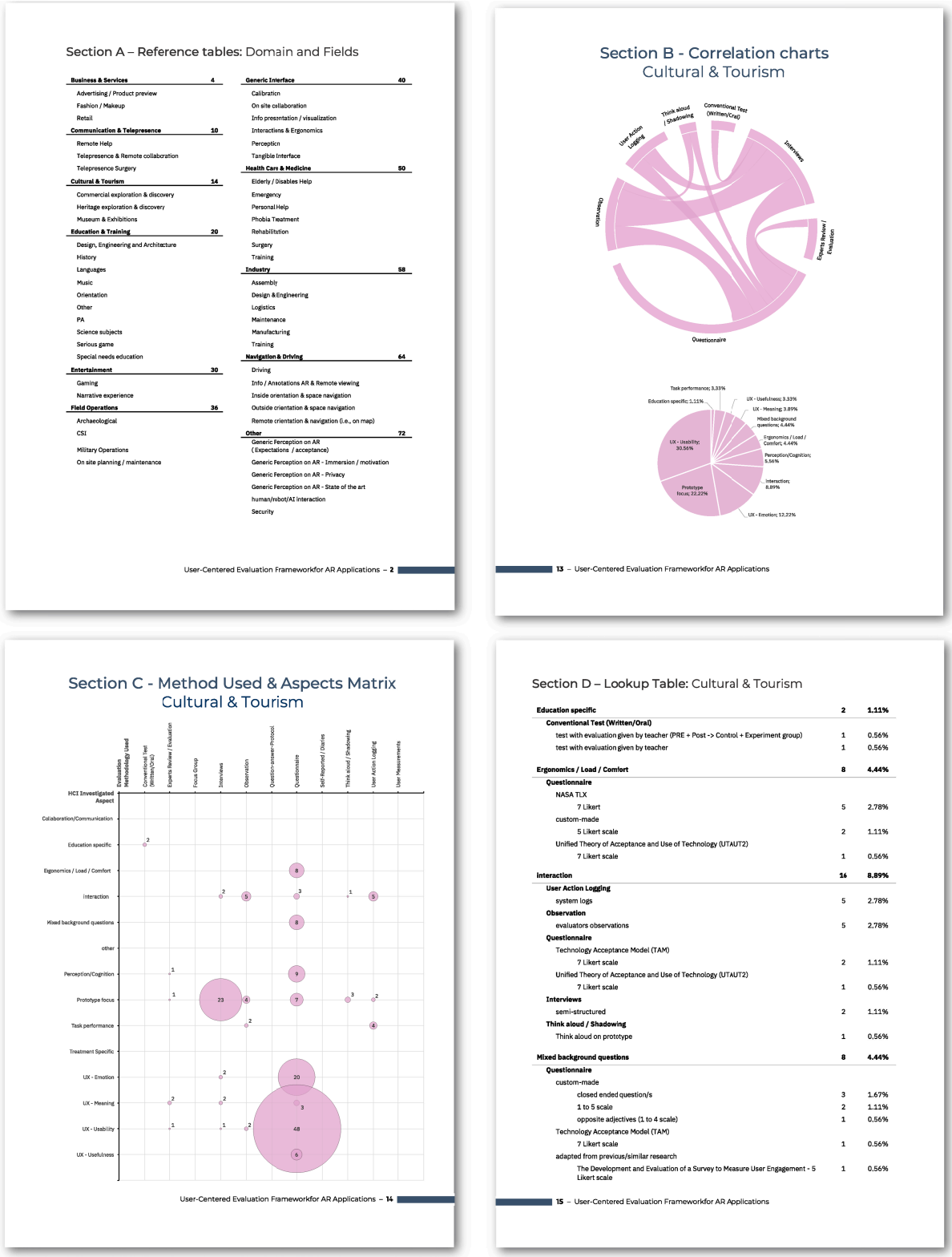
This choice was made to keep the framework simply organized and readable. The framework includes the results of the paper analysis organized into four main sections, as shown in Figure 1:

- A. Reference tables better illustrate the differences between each domain, evaluation methods, and HCI aspects. They aim to describe the main differences between the various categories to allow for a more accessible selection.
- B. Correlation charts: graphs that show the correlation between the methods used and the most investigated HCI aspects divided into different domains. The metrology correlation graphs reveal which methods are frequently used together, and we can see the strength of each link. The pie chart shows the percentage of studies investigating each HCI aspect.
- C. Method Used and Aspects Matrix: This bubble graph correlates the method used for each HCI aspect for each domain. By looking at the size of each bubble, we can easily see if one pair is used more than the other.
- D. Lookup tables: Tables that report all the evaluation methods used, subdivided by the investigated HCI aspect and our subdivision of user evaluation methods.

To use the framework, the evaluators should follow the procedure depicted in Figure 2 during the evaluation design phase. This procedure is inspired by the evaluation procedures defined in [31,32].

1. The researcher chooses the most appropriate application domain for their AR system (Section A).
2. Using the domain chart in Section B, the researcher should decide what type of HCI investigations they want to conduct. This section can be used to evaluate past explorations and determine the need for further investigation.
3. The domain matrix in Section C displays which method was most frequently used in the past based on the selected HCI aspect. It considers the various possible methods.
4. Section D provides the lookup table of the correspondent domain and allows a more detailed evaluation of the possible tools and their subtypes.
5. Finally, the correlation chart in Section B could be used to refine the selection of the evaluation method and better cover the different HCI aspects of the evaluated AR system.

It is worth noting that the framework could propose irrelevant evaluation methods if no studies have been conducted in a specific sector or the investigated HCI aspect has not yet been explored. In this case, the researcher could consider the generic interfaces domain.



Section A – Reference tables: Domain and Fields

Business & Services	**4**
Advertising / Product preview	
Fashion / Makeup	
Retail	
Communication & Telepresence	**10**
Remote Help	
Telepresence & Remote collaboration	
Telepresence Surgery	
Cultural & Tourism	**14**
Commercial exploration & discovery	
Heritage exploration & discovery	
Museum & Exhibitions	
Education & Training	**20**
Design, Engineering and Architecture	
History	
Languages	
Music	
Orientation	
Other	
PA	
Science subjects	
Serious game	
Special needs education	
Entertainment	**30**
Gaming	
Narrative experience	
Field Operations	**36**
Archaeological	
CSI	
Military Operations	
On site planning / maintenance	
Generic Interface	**40**
Calibration	
On site collaboration	
Info presentation / visualization	
Interactions & Ergonomics	
Perception	
Tangible Interface	
Health Care & Medicine	**50**
Elderly / Disables Help	
Emergency	
Personal Help	
Phobia Treatment	
Rehabilitation	
Surgery	
Training	
Industry	**58**
Assembly	
Design & Engineering	
Logistics	
Maintenance	
Manufacturing	
Training	
Navigation & Driving	**64**
Driving	
Info / Anotations AR & Remote viewing	
Inside orientation & space navigation	
Outside orientation & space navigation	
Remote orientation & navigation (i.e., on map)	
Other	**72**
Generic Perception on AR (Expectations / acceptance)	
Generic Perception on AR - Immersion / motivation	
Generic Perception on AR - Privacy	
Generic Perception on AR - State of the art	
human/robot/AI interaction	
Security	

User-Centered Evaluation Frameworkfor AR Applications – **2**

Section B - Correlation charts
Cultural & Tourism

13 – User-Centered Evaluation Frameworkfor AR Applications

Section C - Method Used & Aspects Matrix
Cultural & Tourism

User-Centered Evaluation Frameworkfor AR Applications – **14**

Section D – Lookup Table: Cultural & Tourism

Education specific	**2**	**1.11%**
Conventional Test (Written/Oral)		
test with evaluation given by teacher (PRE + Post -> Control + Experiment group)	1	0.56%
test with evaluation given by teacher	1	0.56%
Ergonomics / Load / Comfort	**8**	**4.44%**
Questionnaire		
NASA TLX		
7 Likert	5	2.78%
custom-made		
5 Likert scale	2	1.11%
Unified Theory of Acceptance and Use of Technology (UTAUT2)		
7 Likert scale	1	0.56%
Interaction	**16**	**8.89%**
User Action Logging		
system logs	5	2.78%
Observation		
evaluators observations	5	2.78%
Questionnaire		
Technology Acceptance Model (TAM)		
7 Likert scale	2	1.11%
Unified Theory of Acceptance and Use of Technology (UTAUT2)		
7 Likert scale	1	0.56%
Interviews		
semi-structured	2	1.11%
Think aloud / Shadowing		
Think aloud on prototype	1	0.56%
Mixed background questions	**8**	**4.44%**
Questionnaire		
custom-made		
closed ended question/s	3	1.67%
1 to 5 scale	2	1.11%
opposite adjectives (1 to 4 scale)	1	0.56%
Technology Acceptance Model (TAM)		
7 Likert scale	1	0.56%
adapted from previous/similar research		
The Development and Evaluation of a Survey to Measure User Engagement - 5 Likert scale	1	0.56%

15 – User-Centered Evaluation Frameworkfor AR Applications

Figure 1. The four sections of the framework are extracted from one of the identified domains (e.g., Cultural and Tourism).

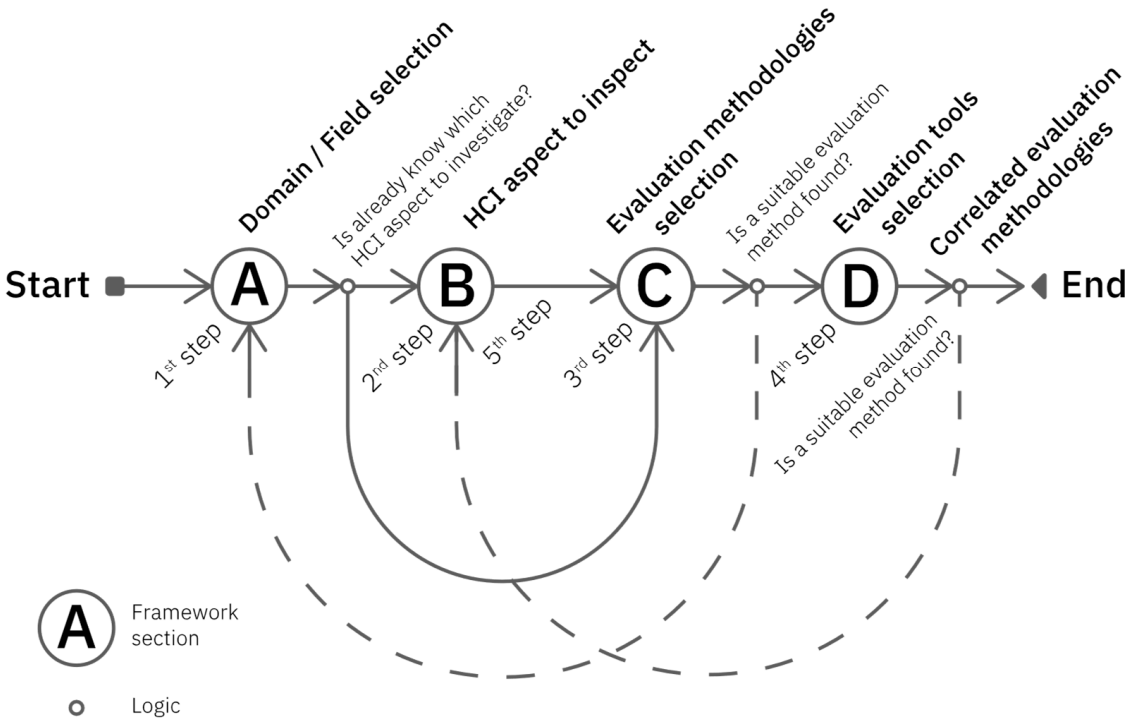


Figure 2. Framework usage process steps.

6. Discussion

Although the framework proposed in this study provides several outcomes, some limitations must be addressed to improve its reliability and usability across a broader range of use cases. This is particularly evident for evaluations carried out in domains where the available study samples are relatively small, and the framework may not always fully satisfy the researchers' needs. However, in a recent review [33], the authors confirm that just a small portion of studies include rigorous evaluation activities with users, and often, the evaluation methods are not reported. In addition, [34,35] highlighted how, despite the advancements in AR research, there is still a lack of well-defined and standardized user evaluation methods.

This is mainly due to the paper sourcing phase, which was limited to collecting papers that had already been partially cataloged in previous surveys. Indeed, as pointed out in [8], concentrating solely on a few sources for analysis and scope may exclude potentially impactful research while disregarding influential research from other fields that explore AR's potential uses. To cope with this problem, a more extensive and systematic survey of the primary online databases (such as Scopus [36], Science Direct [37], IEEE Xplore Digital Library [38], Research Gate [39], and Google Scholars [40]) in the domain of HCI, UX, UI, usability, and user evaluations in the field of AR technologies is needed.

This problem stems from the quantity and quality of data collected, resulting in a lack of potential framework outcomes. Many sources mainly refer to one specific domain or research field, where the analysis is focused on AR learning applications, and it can constitute a bias toward other types of evaluations and HCI aspects. This is still due to the collections of papers that feed the framework. These collections did not use the same acceptance criteria during their evaluation's selection phases, and more importantly, their material classification does not entirely align with the one we proposed. This discrepancy creates a bias and does not represent AR evaluation's current state of the art. An example of this can be found in [28]. Our main objective was to identify the tools and methods used to evaluate AR technologies in the context of education and training. Upon conducting the cluster analysis, we also decided to include a publication in our papers sample, as many documents discussed topics beyond education and trading, including navigation, general

interfaces, and more. However, this could suggest that our sample may not represent a comprehensive overview of the latest advancements in the field, making it difficult to understand the domain.

Another substantial limitation of this work was performing a manual analysis of the paper, which is inevitably slow and prone to errors and misjudgments of the materials investigated. To speed up this process and thus guarantee a more vast and reliable inquiry of the materials, other tools could be employed, such as Large Language Models (LLM) [41], which recently have become readily available and widely adopted and that can be trained on the material already collected, to then be used for other collections to extend the current framework in a programmatic and fast way.

7. Conclusions

This work has provided a comprehensive examination of the evaluation methods employed within the field of AR, with a particular emphasis on selecting the most appropriate methods for assessing these evolving systems and interfaces. Despite over three decades of active research in AR, numerous challenges persist, both technically and from a user perspective. While a significant effort has been dedicated to addressing technical difficulties, there has been a notable need for research focused on the user-centric aspects of AR. This oversight has been identified as a critical element in advancing AR technology. The success of AR devices hinges on a concerted shift towards user-centered, design-oriented research aimed at crafting well-designed user experiences.

The absence of formal user evaluations can be attributed, in part, to a lack of understanding among researchers regarding which evaluation methods are best suited to their specific circumstances, given the vast and fragmented nature of the AR field. Therefore, there is a need to reassess and re-evaluate the types of evaluations conducted thus far to establish a robust foundation for the future assessments of AR systems, with a persistent focus on the end-users and their diverse needs, emotional considerations, and desires factors, often overshadowed by the technology-driven agenda prevalent in the AR domain.

The practical framework developed in this study guides the selection of optimal user evaluation methods for AR systems. A systematic approach was employed to identify and categorize relevant studies, creating a comprehensive framework organized around three main pillars: application domains, evaluation methods, and the investigated HCI aspects. This work analyzed 473 individual studies from 2001 to 2018 and developed a framework that provides valuable insights for researchers and practitioners. Moving forward, the AR community must continue to prioritize user-centric research, leveraging frameworks like the one proposed herein to drive AR technology advancements that prioritize usability, user satisfaction, and overall user experience.

In the future, we have plans to develop an interactive online platform that will allow researchers to access data quickly, easily, and efficiently. This platform could also facilitate sharing AR evaluations, including their evaluation methods and HCI aspects, thereby expanding the framework data and making them more usable. Additionally, this platform could serve as a repository for studies used in AR system evaluations, and a meeting point for researchers required to conduct assessments with real users.

We want to emphasize that this framework should be considered an ongoing project. It will continue to evolve and improve over time to provide the best possible tools for researchers. As more researchers use this framework, they will contribute to its growth and development, making it more mature and widely adopted in everyday life to achieve the so-called "next-generation interface" [4].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/mti8050041/s1.

Author Contributions: Conceptualization, G.C.; methodology, A.P.; software, A.P.; validation, A.P., G.C.; formal analysis, A.P. and G.C.; investigation, A.P.; data curation, A.P.; writing—original draft

preparation, A.P.; writing—review and editing, G.C.; visualization, A.P.; supervision, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the data is contained within the article and the Supplementary Materials.

Acknowledgments: This research was supported by the i.Drive Lab (http://www.idrive.polimi.it/, accessed on 10 April 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Evangelista, A.; Ardito, L.; Boccaccio, A.; Fiorentino, M.; Messeni Petruzzelli, A.; Uva, A.E. Unveiling the Technological Trends of Augmented Reality: A Patent Analysis. *Comput. Ind.* **2020**, *118*, 103221. [CrossRef]
2. Mackay, W.E. Augmented Reality: Linking Real and Virtual Worlds: A New Paradigm for Interacting with Computers. In Proceedings of the Working Conference on Advanced Visual Interfaces, L'Aquila, Italy, 24–27 May 1998; ACM: New York, NY, USA, 1998; pp. 13–21.
3. Wellner, P.; Mackay, W.; Gold, R. Back to the Real World. *Commun. ACM* **1993**, *36*, 24–26. [CrossRef]
4. Jacob, R.J.K. What Is the next Generation of Human-Computer Interaction? In Proceedings of the CHI '06 Extended Abstracts on Human Factors in Computing Systems, Montréal, QC, Canada, 21–27 April 2006; ACM: New York, NY, USA, 2006; pp. 1707–1710.
5. Dünser, A.; Grasset, R.; Hartmut, S.; Billinghurst, M. *Applying HCI Principles to AR Systems Design*; University of Canterbury: Christchurch, New Zealand, 2007.
6. Swan, J.E., II; Gabbard, J.L. Survey of User-Based Experimentation in Augmented Reality. In Proceedings of the 1st International Conference on Virtual Reality, Las Vegas, NY, USA, 22–27 July 2005; pp. 1–9.
7. Dey, A.; Billinghurst, M.; Lindeman, R.W.; Swan, J.E. A Systematic Review of 10 Years of Augmented Reality Usability Studies: 2005 to 2014. *Front. Robot. AI* **2018**, *5*, 37. [CrossRef] [PubMed]
8. Kim, K.; Billinghurst, M.; Bruder, G.; Duh, H.B.-L.; Welch, G.F. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 2947–2962. [CrossRef] [PubMed]
9. Mautino, S.; Melnykowycz, M. User-Experience in Wearable Displays: A Proposal for Standards Definition—The Test Case of Immersive Experiencing for User Engagement in Story Telling Applications. 1 March 2013. Available online: https://www.researchgate.net/profile/Sara-Mautino/publication/236133788_User-experience_in_wearable_displays_a_proposal_for_standards_definition_-_The_test_case_of_immersive_experiencing_for_user_engagement_in_story_telling_applications_-/links/5b22e1caaca272277fb03ee3/User-experience-in-wearable-displays-a-proposal-for-standards-definition-The-test-case-of-immersive-experiencing-for-user-engagement-in-story-telling-applications.pdf (accessed on 16 March 2024).
10. Sharp, H.; Preece, J.; Rogers, Y. *Interaction Design: Beyond Human-Computer Interaction*, 5th ed.; Wiley: Indianapolis, IN, USA, 2019; ISBN 978-1-119-54725-9.
11. Nielsen, J. What Is Usability? In *User Experience Re-Mastered*; Elsevier: Amsterdam, The Netherlands, 2010; pp. 3–22; ISBN 978-0-12-375114-0.
12. Nielsen, J.; Molich, R. Heuristic Evaluation of User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, 1–5 April 1990; Association for Computing Machinery: New York, NY, USA, 1990; pp. 249–256.
13. Molin, L. Wizard-of-Oz Prototyping for Co-Operative Interaction Design of Graphical User Interfaces. In Proceedings of the Third Nordic Conference on Human-Computer Interaction, Tampere, Finland, 23–27 October 2004; Association for Computing Machinery: New York, NY, USA, 2004; pp. 425–428.
14. Dünser, A.; Billinghurst, M. Evaluating Augmented Reality Systems. In *Handbook of Augmented Reality*; Furht, B., Ed.; Springer: New York, NY, 2011; pp. 289–307; ISBN 978-1-4614-0063-9.
15. Greenberg, S.; Buxton, B. Usability Evaluation Considered Harmful (Some of the Time). In Proceedings of the CHI 2008 Proceedings, Florence, Italy, 5 April 2008.
16. Dünser, A.; Grasset, R.; Billinghurst, M. A Survey of Evaluation Techniques Used in Augmented Reality Studies. In Proceedings of the ACM SIGGRAPH ASIA 2008 Courses on—SIGGRAPH Asia '08, Singapore, 10–13 December 2008; ACM Press: Singapore, 2008; pp. 1–27.
17. Bai, Z.; Blackwell, A.F. Analytic Review of Usability Evaluation in ISMAR. *Interact. Comput.* **2012**, *24*, 450–460. [CrossRef]
18. Lew, H.L.; Poole, J.H.; Lee, E.H.; Jaffe, D.L.; Huang, H.-C.; Brodd, E. Predictive Validity of Driving-Simulator Assessments Following Traumatic Brain Injury: A Preliminary Study. *Brain Inj.* **2005**, *19*, 177–188. [CrossRef] [PubMed]
19. *ISO 9241-210:2019*; Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems. International Organization for Standardization (ISO): Geneva, Switzerland, 2019; p. 33.

20. Cavalcanti, V.C.; de Santana, M.I.; Gama, A.E.F.D.; Correia, W.F.M.; Arya, A. Usability Assessments for Augmented Reality Motor Rehabilitation Solutions: A Systematic Review. *Int. J. Comput. Games Technol.* **2018**, *2018*, 5387896. [CrossRef]
21. Zarour, M.; Alharbi, M. User Experience Framework That Combines Aspects, Dimensions, and Measurement Methods. *Cogent Eng.* **2017**, *4*, 1421006. [CrossRef]
22. Irshad, S.; Rambli, D.R.A. Preliminary User Experience Framework for Designing Mobile Augmented Reality Technologies. In Proceedings of the 2015 4th International Conference on Interactive Digital Media (ICIDM), Bandung, Indonesia, 1–5 December 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–4.
23. Arfini, S.; Bellani, P.; Picardi, A.; Yan, M.; Fossa, F.; Caruso, G. Design for Inclusivity in Driving Automation: Theoretical and Practical Challenges to Human-Machine Interactions and Interface Design. In *Connected and Automated Vehicles: Integrating Engineering and Ethics*; Fossa, F., Cheli, F., Eds.; Studies in Applied Philosophy, Epistemology and Rational Ethics; Springer Nature: Cham, Switzerland, 2023; Volume 67, pp. 63–85; ISBN 978-3-031-39990-9.
24. Preece, J.; Rogers, Y.; Sharp, H. What Is Interaction Design? In *Interaction Design: Beyond Human-Computer Interaction*; Wiley: Chichester, UK, 2015; ISBN 978-1-119-08879-0.
25. Hartson, R.; Pyla, P.S. What Are UX and UX Design? In *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*; Morgan Kaufmann: Amsterdam, The Netherlands, 2012; ISBN 978-0-12-385242-7.
26. Lallemand, C.; Gronier, G.; Koenig, V. User Experience: A Concept without Consensus? Exploring Practitioners' Perspectives through an International Survey. *Comput. Hum. Behav.* **2015**, *43*, 35–48. [CrossRef]
27. Kostaras, N.; Xenos, M. Usability Evaluation of Augmented Reality Systems. *Intell. Decis. Technol.* **2012**, *6*, 139–149. [CrossRef]
28. Lim, K.C.; Selamat, A.; Alias, R.A.; Krejcar, O.; Fujita, H. Usability Measures in Mobile-Based Augmented Reality Learning Applications: A Systematic Review. *Appl. Sci.* **2019**, *9*, 2718. [CrossRef]
29. Bach, C.; Scapin, D. Obstacles and Perspectives for Evaluating Mixed Reality Usability. 1 January 2004. Available online: https://www.researchgate.net/publication/221104007_Obstacles_and_Perspectives_for_Evaluating_Mixed_Reality_Usability (accessed on 16 March 2024).
30. Kostaras, N.; Xenos, M. Assessing the Usability of Augmented Reality Systems. In Proceedings of the 13th Panhellenic Conference on Informatics, Corfu, Greece, 10 September 2009; pp. 197–201.
31. Gabbard, J.L.; Hix, D.; Swan, J.E. User-Centered Design and Evaluation of Virtual Environments. *IEEE Comput. Graph. Appl.* **1999**, *19*, 51–59. [CrossRef]
32. Gabbard, J.L.; Swan, J.E. Usability Engineering for Augmented Reality: Employing User-Based Studies to Inform Design. *IEEE Trans. Vis. Comput. Graph.* **2008**, *14*, 513–525. [CrossRef] [PubMed]
33. Cosio, L.D.; Buruk, O.O.; Fernández Galeote, D.; Bosman, I.D.V.; Hamari, J. Virtual and Augmented Reality for Environmental Sustainability: A Systematic Review. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–23.
34. Al-Ansi, A.M.; Jaboob, M.; Garad, A.; Al-Ansi, A. Analyzing Augmented Reality (AR) and Virtual Reality (VR) Recent Development in Education. *Soc. Sci. Humanit. Open* **2023**, *8*, 100532. [CrossRef]
35. Massa, E.; Ladhari, R. Augmented Reality in Marketing: Conceptualization and Systematic Review. *Int. J. Consum. Stud.* **2023**, *47*, 2335–2366. [CrossRef]
36. Scopus—Document Search. Available online: https://www.scopus.com/search/form.uri?display=basic#basic (accessed on 16 March 2024).
37. ScienceDirect.Com | Science, Health and Medical Journals, Full Text Articles and Books. Available online: https://www.sciencedirect.com/ (accessed on 16 March 2024).
38. IEEE Xplore. Available online: https://ieeexplore.ieee.org/Xplore/home.jsp (accessed on 16 March 2024).
39. ResearchGate | Find and Share Research. Available online: https://www.researchgate.net/ (accessed on 16 March 2024).
40. Google Scholar. Available online: https://scholar.google.com/ (accessed on 16 March 2024).
41. Mimno, D.; Wallach, H.M.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 262–272.