MDPI

Systematic Review

Assessing Human Factors in Virtual Reality Environments for Industry 5.0: A Comprehensive Review of Factors, Metrics, Techniques, and Future Opportunities

Oscar Escallada *[ID], Ganix Lasa [ID], Maitane Mazmela [ID], Ainhoa Apraiz [ID], Nagore Osa [ID] and Hien Nguyen Ngoc

Mondragon Unibertsitatea, Faculty of Engineering, Mechanics and Industrial Production,
Design Innovation Centre, Loramendi, 4, 20500 Mondragon, Spain; glasa@mondragon.edu (G.L.);
mmazmela@mondragon.edu (M.M.); aapraizi@mondragon.edu (A.A.); nosa@mondragon.edu (N.O.);
n.nnhien1990@gmail.com (H.N.N.)
*   Correspondence: oescallada@mondragon.edu

Abstract: Industry 5.0, the latest evolution in industrial processes, builds upon the principles of Industry 4.0 by emphasizing human-centric approaches and the integration of virtual reality technologies. This paradigm shift underscores the importance of collaboration between humans and advanced technologies with a focus on optimizing efficiency, safety, and worker skill development. Based on the PRISMA 2020 guidelines, this study conducts a systematic literature review, identifying 328 papers from databases. After applying inclusion and exclusion criteria, 24 papers were selected for detailed analysis. The review provides valuable insights into the diverse evaluation methods employed in the literature, and a detailed classification of 29 human factors with their associated metrics. Despite the absence of a standardized method for assessing human factors in VR experiences, this comprehensive analysis of 240 different ways of measuring factors highlights the current state of evaluating human-centered VR experiences in Industry 5.0. While the review reveals some limitations such as potential bias in study selection and heterogeneity of methods, it also identifies significant research gaps and proposes future directions. This study contributes to the establishment of a coherent structure for future research and development in human-centered design within the rapidly evolving landscape of Industry 5.0, paving the way for more effective and standardized approaches in the future.

Keywords: human factors; virtual reality; industry 5.0; interaction; human-centered design

1. Introduction

As the transition from one industrial era to the next looms, the emergence of Industry 5.0 emerges to address the limitations of its predecessor while introducing new opportunities and challenges. This revolution envisions a landscape of autonomous manufacturing enhanced with human intelligence: a concept articulated by Mourtzis et al. [1] and Schroeter [2]. This fusion of automation and human intelligence redefines production, creating a collaborative ecosystem where machines and humans synergistically enhance productivity and innovation.

Within this shifting panorama, Industry 5.0 emerges as a beacon of hope, addressing the limitations and complexities posed by its predecessor. Its vision extends to a more inclusive and sustainable production model, one that values not only technical efficiency but also the well-being and growth of the workforce. This human-centric approach is

vividly illustrated by Industry 5.0's aspiration to harness the potential of innovative digital technologies while nurturing a harmonious human-machine interaction [3,4]. This interaction envisions smart industrial environments where human operators and machines collaborate seamlessly, yielding a new era of production excellence [5].

In alignment with Industry 5.0 principles, virtual reality (VR) training systems represent innovative solutions that blend technological advancements with human-centric design. This revolution of VR was catalyzed by the accessibility of immersive head-mounted displays (HMDs) like the Meta Quest, HTC Vive, and Sony PlayStation VR [6], which spurred its widespread adoption across various domains such as medicine, education, and engineering. Nevertheless, beyond its technological aspects, VR's integration necessitates a profound understanding of human–computer interaction and user experience [7,8]. Crafting a superlative user experience involves factors like presence, immersion, and engagement, which elevate the user's sense of interaction [9]. Additionally, reducing VR-induced symptoms and effects (VRISEs) becomes paramount in ensuring an enriching and holistic User Experience (UX) [10].

In the field of VR training systems, the landscape is dynamic, yet full of intricate challenges. As the popularity of VR surges [11] it becomes evident that the effectiveness and efficiency of these technologies as training mediums are marked by diverse outcomes. Daling and Schlittmeier [12] emphasize the complexity of this research field, pointing to an excess of results that are often heterogeneous and even contradictory. While the potential of VR is undeniable, considerable systematic research is required to fully harness their capabilities [13]. However, such research must consider multiple human factor issues for these VR systems to be effective and well-received by users [13].

The diversity of research methodologies and objective performance measures further complicates the comparability of results [12]. This variance is reflective of the intricate interplay between the multifaceted factors involved in VR experiences. Kaplan et al. [14] illuminate the difficulties in achieving consistent findings, attributing them to the variability in MR technologies, task specifics, training methods, and performance metrics. The paucity of data makes the conduct of meta-analyses a challenging endeavor [14], underscoring the need for comprehensive research that encompasses a broader range of outcome variables.

The synthesis of these insights paints a comprehensive picture of the challenges and opportunities presented by VR training systems. While their potential is vast, their implementation requires a deep understanding of human factors, design principles, and a commitment to systematic research. As these technologies continue to evolve, researchers remain engaged in the pursuit of optimized training paradigms that seamlessly merge technology and human experience.

2. Materials and Methods

This section outlines the framework for investigating the broader implications of VR in industrial applications, with a particular focus on evaluating and understanding the impact of human factors. The increasing complexity and sophistication of VR technologies in industrial settings highlight the necessity for a systematic literature review (SLR) to comprehensively gather, evaluate, and synthesize existing research findings. The diversity in outcomes and methodologies from prior studies emphasizes the need for a systematic approach to amalgamate data, identify consistent trends, and address discrepancies effectively. Furthermore, the rapid advancements in VR technologies necessitate including the most recent studies to fully assess their potential impact on user experience and performance within industrial environments.

This SLR will employ rigorous criteria to select studies, focusing on those that illuminate the integration of human factors within VR systems and their effects on user interaction

and system efficacy. By consolidating and synthesizing the wide range of existing research, this review aims to provide a clearer understanding of the current landscape and highlight areas needing further investigation, thus paving the way for optimized application of VR in industrial contexts. In this literature review method, the guidelines proposed by Kitchenham [15], as well as the tools suggested by Carrera et al. [16] and PRISMA [17] were used to carry out the SLR.

According to Kitchenham [15], the SLR has three phases: planning, conducting, and documenting. These phases have their own components that are showed in Figure 1 and detailed in the following subsections. In addition, Carrera et al. [16] propose the use of Parsif.al [18] in some of these components. This tool allows us to define goals and objectives, import articles using BibTeX files, eliminate duplicates, define selection criteria, and generate reports. This tool has been used in the steps of the research strategy, publication selection and publication quality assessment as this tool helps to organize all the papers through these stages.
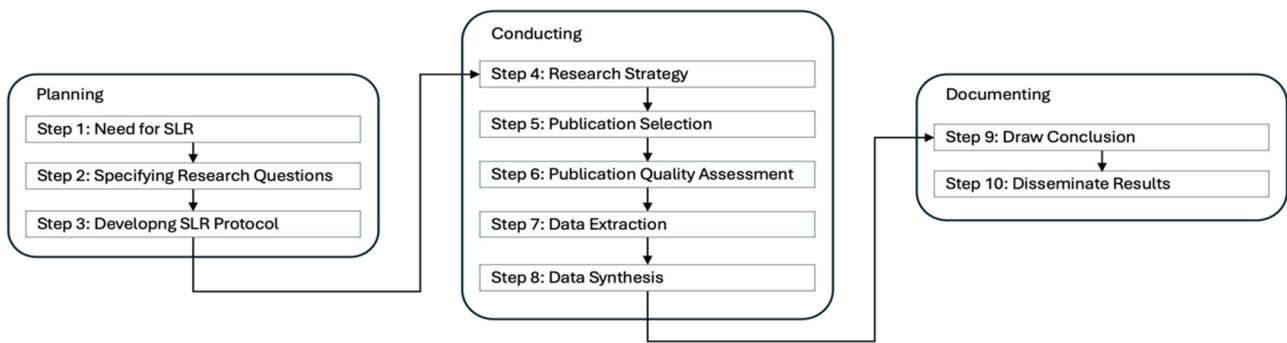

Figure 1. Stages of a systematic literature review. Adapted from Kichenham [15].

2.1. Planning and Review

The planning phase of an SLR begins with recognizing the need for such a review and results in developing the review protocol.

2.1.1. Need for SLR

In the planning stage of an SLR, the initial step involves determining the necessity for conducting an SLR [16]. The need for the SLR in this research is detailed in Section 1. Additionally, the broad objectives and scope of the study are delineated using the population, intervention, comparison, outcome, and context (PICOC) framework as outlined in Table 1.

Table 1. PICOC criteria.

Criteria	RQ
Population	Industrial workers engaging with VR technologies in their operational settings.
Intervention	Implementation and usage of VR technologies aimed at enhancing human-centric approaches in Industry 5.0.
Comparison	Analysis of different human factors evaluation methods and their characteristics for assessing VR technologies.
Outcome	Evaluation of human factors in VR environments. Classification of the methods tools and measurements
Context	Industrial settings where VR technologies are integrated, such as manufacturing plants and engineering firms, but also laboratory tests and experiments.

2.1.2. Specifying Research Questions

In the context of this systematic literature review (SLR), the investigation was guided by two different research questions. The comprehensive resolution of both is imperative for the successful execution of this systematic review. These research questions are shown in Table 2.

Table 2. Definition of research questions.

ID	Research Questions
RQ1	Is there a model for evaluating industrial virtual reality experiences that includes human factors?
RQ2	What human factors are measured in industrial virtual reality experience evaluation, and how are they assessed?

The motivations behind the research questions in this study are already mentioned, but in summary is to focus on enhancing the understanding and evaluation of human factors within industrial VR environments. For RQ1, the primary motivation is to determine whether existing models can effectively incorporate human factors when assessing VR experiences in industrial settings. This inquiry aims to facilitate improvements in user interaction and safety (among other factors) by identifying models that holistically evaluate human contributions to VR applications. Regarding RQ2, the motivation is to identify the human factors that most significantly influence VR experiences in industrial contexts and to understand how these factors are currently assessed. This is aimed at enhancing the design and implementation of VR systems by ensuring that the most impactful human factors are accurately evaluated and addressed. Together, these motivations drive the research towards developing a deeper comprehension of how human elements interact with VR technologies, potentially leading to more effective and human-centric VR systems in industrial applications.

2.1.3. Developing SLR Protocol

A pre-defined SLR protocol is necessary, which specifies methods that will be used to conduct a specific literature review and will reduce researcher bias [15]. A review protocol was developed by a teamwork of authors and externally evaluated by an expert having experience in SLRs before its execution [16]. The developed protocol is based on PRISMA protocol's guidelines [17].

2.2. Conducting the Review

Following with the second phase of SLRs, a definition of the search strategy is needed. In this section, this strategy, plus the publication selection, the quality assessment as well as the data extraction and its posterior synthesis are described.

2.2.1. Search Strategy

To perform the systematic review, different terms have been identified. They can be classified into these four groups:

- Virtual reality/VR.
- Industry 4.0/Industry 5.0/operator/manufacturing.
- Human factors/cognition/cognitive/user experience/UX/interaction/interactive.
- Evaluation/assessment.

Thus, the following main search equation has been created: (TITLE-ABS-KEY ({Virtual Reality} OR vr) AND TITLE-ABS-KEY ({industry 4.0} OR {industry 5.0} OR operator OR manufacturing) AND TITLE-ABS-KEY ({human facto*} OR {cogniti*} OR {User Experience} OR ux OR interacti*) AND TITLE-ABS-KEY (evaluation OR assessment)).

The electronic databases used for the search are the following: EI Compendex, IEEE Digital Library, INSPEC, ISI Web of Science, Science Direct, and Scopus. Considering the general equation mentioned above, it has been necessary to adapt it to the different databases because each one has different search methods. Nevertheless, there are general criteria that must be met by all the articles selected. First, they must be peer-reviewed journal articles. A 10-year age period has been determined for the articles, which means that documents published between 2013 and June 2023 will be accepted. Regarding the decision to limit the filtering of the papers to a period of 10 years, it has been made especially because the current VR devices on the market are relatively new due to the rapid advancement of technology. Hence, also the human factors interest related to them is recent. In terms of the language, all papers that are not written in English or Spanish will be excluded.

Considering all these characteristics, a total of 588 papers have been identified. The vast majority were found in Scopus (230) and in EI Compendex (146).

2.2.2. Publication Selection

In this next step, we proceeded to review the literature. First, we defined inclusion and exclusion criteria for the articles to ensure that the selected studies were related to the previously defined topic. Then, we proceeded to conduct a QA to identify the most relevant articles.

Once the identification of the 588 papers is complete, different criteria have been defined for the inclusion or exclusion of these. Apart from the language limitation, that only papers written in English or Spanish will be considered due to the authors' language understanding limitations, it has also been considered if the paper appears more than once in the search, that is, if it is duplicated. A total of 260 duplicate papers were identified. Finally, two other inclusion/exclusion criteria were considered. To ensure the credibility of the published papers, we excluded papers that were not peer-reviewed and added the condition that they were published in journals (excluding proceedings, conferences, and the like). The inclusion and exclusion criteria are shown in Table 3.

Table 3. Inclusion and exclusion criteria for paper evaluation.

	Inclusion Criteria	**Exclusion Criteria**
Relationship with the topic	The paper responds to at least one of the two research questions.	The paper does NOT respond to any research question.
Language	The full text is written in English or Spanish.	The full text is NOT written in English or Spanish.
Duplicated paper	The paper is NOT duplicated in the search.	The paper appears twice or more times as it is duplicated.
Publication	The paper is published as a journal article in the databases studied.	The paper is not peer reviewed or it has been published as proceedings or as a conference paper.

2.2.3. Quality Assessment (QA)

Following the exclusion and inclusion criteria, QA was conducted. This process identified whether the articles were related to the specific topic and whether they were useful in terms of assessing human factors in an industrial VR environment. The papers were divided between two reviewers for evaluation, who then randomly sampled and conducted double evaluations to ensure alignment with the established criteria. Any papers that presented uncertainties were subsequently discussed jointly by the reviewers. Three QA questions were formulated, reviewed, and scored based on the analysis. QA1 is related

to RQ1 and the search of models regarding VR and human factors, while QA2 and QA3 are more related to RQ2:

QA1—The proposed topic is related to human factors in an industrial virtual reality environment? This QA aims to identify papers that at least are focused on human factors in industrial VR environments.

QA2—Does this research help to identify the human factors affecting industrial VR environments? This QA aims to value papers which assist in the identification of human factors influencing industrial VR environments.

QA3—Does this research describe how to evaluate human factors in industrial VR environments? QA3 aims to value papers that show methods, tools, or techniques to evaluate human factors in industrial VR environments.

The objective of this QA was to facilitate the understanding of the studies' appropriateness and usefulness to this current study. For an easy classification, Parsif.al's three-type rating criteria were used: high, medium, and low. A score of 2 was given to studies that fully met the quality standard, a score of 1 was given to studies that partially met the quality standard, and a score of 0 was given to the studies that did not meet the quality standard. Taking this into account, the maximum score for each study is 6 (i.e., $3 \times 2 = 6$), and the lowest possible score is 0 (i.e., $3 \times 0 = 0$).

In this SLR, we considered those articles that obtained a score higher than 3 to ensure high-quality and reliable findings. We finally ended up with a total of 26 articles. Figure 2 shows the process carried out during the literature review. Beginning with a total of 588 papers identified, 260 were discarded for being duplicates. From the 328 left, after reading the titles and abstracts, 259 of these were rejected considering inclusion and exclusion criteria. Finally, the last 69 papers, after reading them one by one, were evaluated taking the QA into account. A total of 24 papers received a score higher than 3 out of 6. These papers were analyzed in depth for data extraction and synthesis.
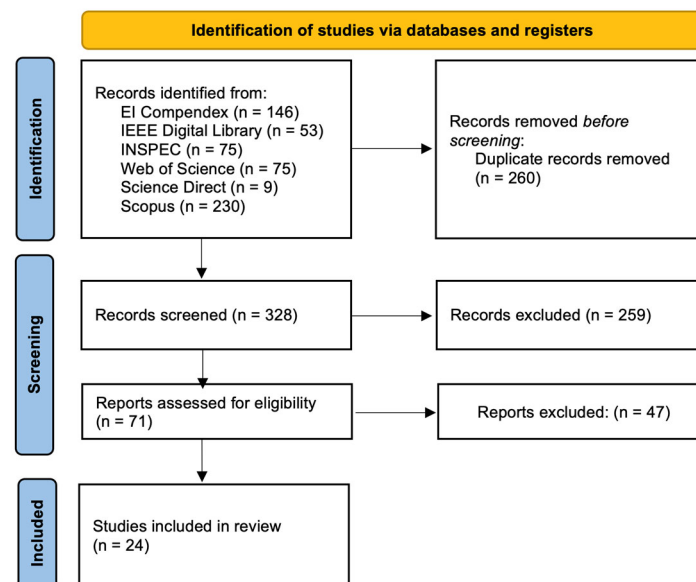
Figure 2. Diagram of the paper filtering process.

2.2.4. Data Extraction

In this step, the 24 selected papers were used for data extraction. The following data were extracted from each study:

The name of authors, title of the paper, name of the journal, impact factor (JCR), publication year, sample of the experiment, gender % of the sample, expertise of the sample (familiarity with the task and with the technology), human factors measured, metrics used,

units of measurement, data collection method, type of VR device used, and extra devices used (biosensors, haptics...).

2.2.5. Data Synthesis

Following the data extraction process, a comprehensive list of human factors, metrics, and techniques has been compiled from the analysis of 24 studies. Additionally, detailed information on the types of VR devices and the auxiliary equipment used to measure human factors was gathered. To ensure a rigorous classification of this data, a peer review of the lists of human factors has been conducted. This step is crucial for validating the accuracy and relevance of the data collected. All this data will be thoroughly described in the subsequent sections.

3. Results

This article evaluates the literature through a critical lens, offering a comprehensive overview of the subject. This section shows the results derived from the characterization of the literature, as well as answers the two research questions set out above.

3.1. Literature Characterization

Prior to discussing the results and analysis of the human factors and addressing the RQs, a brief summary of the general characteristics of the studies involved will be provided.

3.1.1. Evolution in the Field

Considering the years of publication of the articles after removing duplicated papers, one can see the evolution of interest in human factors affecting industrial VR environments. In Figure 3 this interest can be seen more visually. It should be noted that this search was conducted in June 2023 and therefore 2023's data are not displayed on the graph as the year is incomplete. However, it is expected that the number of publications during this year will be even higher as more than 75% of the identified publications date from 2017 to 2023 and VR technologies are increasing even more.
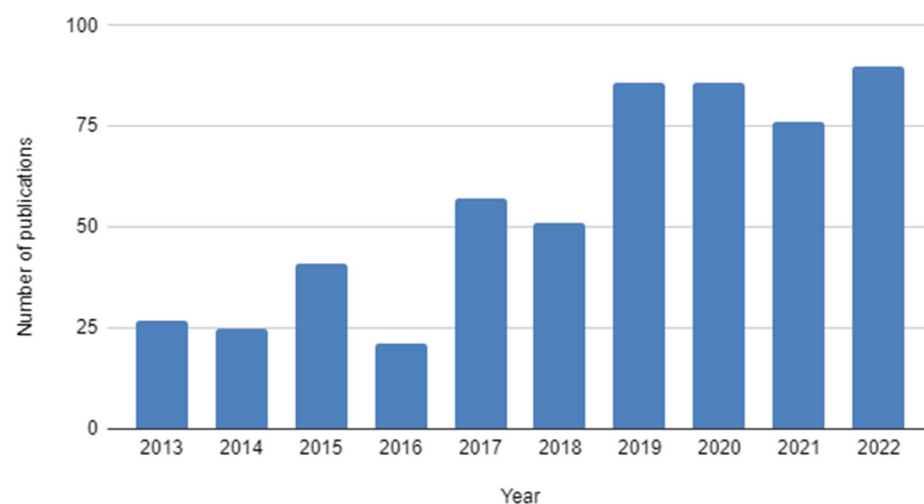

Figure 3. Number of articles identified per year.

3.1.2. Nature of Journals

We analyzed where the 24 selected articles were published and highlighted the wide variety of journals. Only four journals published more than one paper: Applied Sciences MDPI (2), Human Factors (2), Robotics and Computer-Integrated Manufacturing (2), and

VR (2). In terms of indexing, it can be seen in Table 4 that 50% of the identified journals belong to the first quartile, 30% to the second, 5% to the third and 15% to the fourth quartile:

Table 4. Impact of the identified journals (JCR = Journal Citation Reports; nd = not defined).

Journal Title	Quartile	JCR
Applied Sciences MDPI	Q2	2.7
ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems	Q2	3.8
British Journal of Educational Technology	Q1	3.8
Computers in Education Journal	Q4	6.7
DYNA	Q4	3.8
EURASIA Journal of Mathematics Science and Technology Education	Q2	0.903
Human Factors: The Journal of the Human Factors and Ergonomics Society	Q1	3.3
International Journal of Advanced Manufacturing Technology	Q2	3.4
International Journal of Industrial Ergonomics	Q2	3.1
Journal of Cleaner Production	Q1	11.1
Journal of Mechanical Design	Q1	3.3
Manufacturing Technology	Q3	0.9
Multimedia Tools and applications	Q1	3.6
Presence Teleoperators and Virtual Environments	Q4	nd
Production and Manufacturing Research	Q1	4.1
Robotics and Computer-Integrated Manufacturing	Q1	10.4
Robotics MDPI	Q1	3.7
Scientific Reports (Nature)	Q1	4.6
Sustainability MDPI	Q2	3.9
Virtual Reality	Q1	4.2

3.2. RQ-1 Is There a Model for Evaluating Industrial Virtual Reality Experiences That Includes Human Factors?

Among all the papers selected, only 3 papers propose methods for assessing human factors in virtual reality. Peruzzini et al. [19] proposes a method to carry out reliable and effective factory ergonomic analyses during the design stages. It proposes a series of devices as well as seven phases to follow during the process of creating a VM (Virtual Manufacturing) simulation. The phases include creating the virtual scene, configuring the devices, creating the scripts, recording the user's experience and actions, exporting and isolating the most critical operations, and finally, evaluating them after post-processing. It also shows a case study comparing an experience developed using the method and another desktop-based experience. This study also shows the performance indicators measured during the creation of the simulation and the user testing.

The second case study of Peruzzini et al. [19] deals with the UX assessment strategy to identify potentially stressful conditions for workers. For this purpose, different devices are again proposed, including an HMD and different biometric and motion tracking devices to analyze user data. It also discusses the relationships between mental workload, stress, and physical workload, which tools to use and how to calculate the results using the data from the different devices and questionnaires.

In the third case study, Peruzzini et al. [20] defines a multimodal VR set-up for the human-centered design of industrial workstations. In this case it does not focus so much on the process, although it does propose different devices to be used such as eye trackers, motion tracking, etc. The paper focuses more on proposing which factors, metrics, and data collection methods should be collected. Among these factors are efficiency, effectiveness, and satisfaction.

Ahmed and Onan Demirel [21], discuss the challenges of prototyping user-focused emergency situations, noting that most approaches prioritize evaluating human perfor-

mance through usability testing. However, there is currently no guidance on the typology, variety, or complexity of prototypes that should be developed to address users' needs. The aim of their paper is to establish prototyping methodologies for the early design evaluation of human performance. To achieve this, they propose a mixed-prototyping approach that integrates human subjects, CAD, marker-less motion tracking devices, and HMDs, with a specific focus on emergency situations and how to prototype for them effectively. Although there are limited references to models to evaluate industrial VR experiences, including human factors, they collected information on devices, tools, sample sizes, and participant skills from existing studies to identify recurring patterns.

3.2.1. Devices

In terms of devices, two main types of VR devices have been found. CAVE (cave automatic virtual environment) and HMD (head-mounted display). CAVE consists of a room-sized VR system where users stand inside and interact with 3D images projected on its walls using tracking devices, enabling collaborative experiences but at a higher cost and requiring dedicated space. On the other hand, HMDs are wearable devices like goggles or helmets, providing a more personal VR experience with handheld controllers for interaction. While generally designed for individual use, some HMDs support online multiplayer and offer varying cost options, making them more accessible to consumers. The usage rate of HMDs (83%) is significantly higher than CAVE devices (17%). In addition to the core elements to develop these experiences, in most of the studies they have been accompanied by other types of elements mentioned below. Only in Ahmed and Onan Demirel's studies [21] and also Refs. [22–24] are HMDs used alone. The full Table A1 of the devices used can be found in Appendix A.

- Motion capture: Motion capture devices are frequently used. In 10 out of the 24 papers, the use of motion capture devices is mentioned to track the whole body or specific points in more detail. In the case of HMDs, the position of the hands can be tracked, but with devices such as Leap Motion [19,25,26], ART Tracking [27], Empatica E4 [28], etc., much more precise data can be obtained.

- Biosensors: The category of biosensors includes all those devices that allow the collection of physiological data. A review of the literature shows that eight of the twenty articles in which HMDs are used have opted for the use of the HTC Vive Pro Eye [19,22,28–30], which contains a built-in eye tracker. Other researchers use other methods for eye tracking such Tobii glasses [20,31]. In addition to eye tracking, it is common to incorporate heart rate meters [28,30,32], as well as other not so common sensors such as EEGs [33]. In this way, in addition to collecting information through questionnaires and other methods in which participants rate themselves, these biometric data can complement that information from another perspective.

- Sound: Although HMDs have built-in audio, three studies mention the use of headphones or 5.1 external sound systems. Ref. [32], for example, highlights the importance of industrial noise in the environment of the experience, helping acoustically to show the position of hydraulics, or motors in the background. Refs. [29,34] use headphones instead to isolate participants from the outside and create a more immersive experience.

- Controllers: Although the CAVE or HMD itself has its own controllers, two authors mention the use of controllers in the experiments. They refer to non-standard controllers. That is, other than the default ones of the CAVE or HMD itself, considering that they already have them, Refs. [35,36] use a joystick and a keyboard, respectively.

- Physical elements and haptics: The last category of devices is those that in this article have been called physical elements and haptics. The purpose of these elements is to

enhance the experience. For example, Ref. [32] mentions the use of a rowing machine. They overlap the rowing experience with the physical object to replicate the effort and movements. It is similar with Ref. [37], where they recreate the space in which the operator's maintenance space is recreated, with walls to restrict movement and even removable elements such as nuts and bolts. In Ref. [26], they instead use a commercial Moog FCS robot equipped with three cylinders that move the subject who experiences the sensation of being in a hydraulic excavator. Ref. [37] also mentions a robotic arm, this time of a smaller size that allows users to experience a perception of collision between virtual objects.

3.2.2. Tools

In addition to the devices, the different techniques and methods used to collect data from the experiences have been compiled (Figure 4). The most used method is the questionnaire, with standard questionnaires such as NASA-TLX [38], QUIS [39] or UEQ [40], appearing in at least 13 articles. Also, self-generated questionnaires, i.e., questionnaires that are either adaptations of existing questionnaires or created by the researchers of the study, are used in 11 articles. Following this method, the next most used methods were eye-tracking (six items such as Ref. [41]), stopwatch (six) to measure time, and both position tracking (four) and video recordings (four) to make user observations.
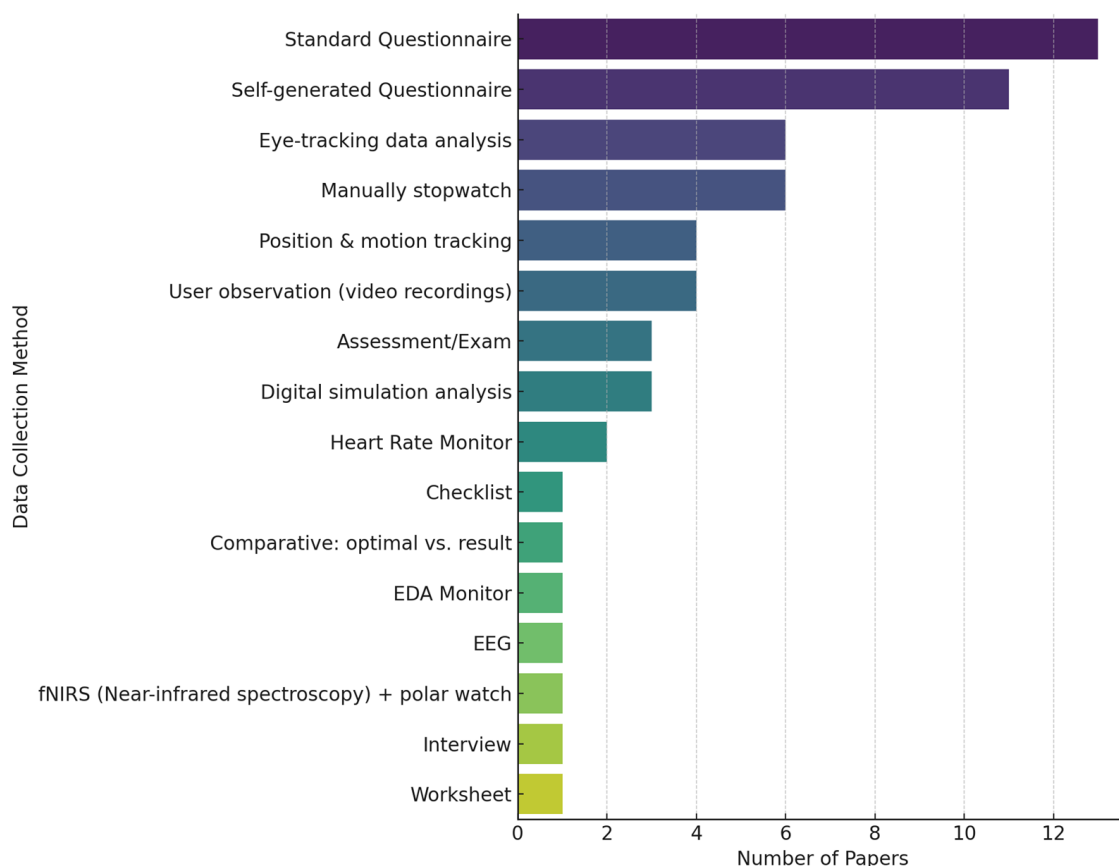
Figure 4. Number of articles per data collection method.

3.2.3. Sample Size and Skills

The sizes of the samples used in the different studies have been collected. It is noticeable the big difference in the number of participants they use for the experiments, varying from one person repeating more than once the experiment, to forty-four participants. Most of the studies justify the number of participants using the ANOVA. The gender of the

participants is mentioned in 16 of the 24 papers. However, it is not considered in the conclusions obtained.

Regarding the skills of the participants, two data have been collected in particular: the experiences the users had in the task of the test, and the skills with XR devices. Not all papers take these data into account. In terms of the topic of the test, for example [32], collects this information prior to the experience to understand if the level of expertise has any effect on the ability to complete the tasks. In the case of skills with XR devices, although not in all cases this information is collected, most of the papers mention prior training or initial contact with the glasses in the case of not having used them before.

3.3. RQ-2 What Human Factors Do They Measure and How Are They Assessed?

By analyzing the literature, a total of 29 human factors have been found to be measured in different ways in the articles. To provide a clear overview of all the factors, a comprehensive Table A2 has been created and is available in the Appendix A. This table offers various classifications of these factors. Firstly, it determines whether the metric used to measure the factor is hedonic or pragmatic. Secondly, it classifies the factors into cognitive, physiological, process-related, or other categories. Additionally, the table maps the unit of measurement, the data collection method, and the author for each factor. This structured approach ensures a thorough and detailed understanding of the data collection methodologies and their respective applications.

Regarding pragmatic and hedonic classification, according to Hassenzahl [42], something may be perceived as pragmatic because it provides effective and efficient means to manipulate the environment. On the contrary, if it is perceived as hedonic, it is because it provides stimulation, identification, or provokes memories. In this case, metrics have been related to the typology of data obtained from users. Both dimensions have been identified as relevant predictors of an interactive product's overall evaluation [43].

In the classification of the three elements, cognitive, physiological, and process, these definitions are defined as follows. In the cognitive domain, metrics are used to evaluate and quantify aspects related to thinking, information processing, and human decision making, and here metrics such as immersion [44] and learnability [23,27] among others, are included. As for the physiological domain, it refers to metrics related to bodily aspects, such as posture and comfort [19], as well as brain activity [33]. Finally, in the process category, metrics related to the development of the activity itself are grouped, such as task execution time [20], task execution accuracy [34], and even error identification or ease of use [24].

To examine the varying significance of the different types of data collection methods, Figure 5 was created. From the table, it is evident that pragmatic methods dominate across most categories, particularly in the process and cognitive dimensions, with 71 and 59 occurrences, respectively, contributing to a substantial total of 177. Hedonic methods appear less frequently, with notable counts only in the cognitive (47) and process (10) categories, resulting in a total of 62. Physiological methods are predominantly pragmatic (47 occurrences) but are scarcely represented in the hedonic (6) category. Overall, the total occurrences of each method type highlights a preference for pragmatic methods in data collection, as shown by the aggregated totals, emphasizing their prominence in this study.

	Cognitive	Process	Physiological	Other	Total
Pragmatic	59	71	47	0	177
Hedonic	46	10	6	0	62
Other	0	0	0	1	1
Total	105	81	53	1	240

Figure 5. Classification of the metrics.

With all this in mind, all these human factors have been classified in a table, indicating the metric with which they have been measured, the unit of measurement, and the technique or method of data collection that has been used. Likewise, these metrics have been classified in two ways. First, they were classified as hedonic or pragmatic. Then, it has been assigned a category among these three: cognitive, physiological, process.

In this exploration, it can be seen how different authors face the assessment. Das et al. [31] initiate the exploration by emphasizing the significance of understanding mental workload and task performance. It echoes the statements of Khamaisi et al. [28], who delves into measuring mental factors to analyze the cognitive workload and mental effort of operators. Both papers underscore the importance of optimizing workload, ultimately enhancing productivity, and preventing health issues in industrial settings. Their shared goal is to ensure worker well-being and satisfaction while supporting human–robot collaboration, thereby reducing industrial costs.

Bu et al. and Morosi and Caruso [26,32] instead align in their focus on user-centered design, performance evaluation, and safety considerations. These papers, like Hoe et al. [44], emphasize user experience, engagement, and satisfaction. By collecting subjective feedback, they aim to iterate design improvements and optimize VR environments for enhanced user performance and satisfaction.

Peruzzini et al. [19] and Rogers et al. [36] take a different but equally crucial path, evaluating ergonomic factors, physical comfort, and safety within industrial contexts. They align with Bernard et al. [37] in their shared pursuit of enhancing operator comfort, reducing the risk of injuries, and minimizing physical and cognitive strain in the workplace.

Both Bernal et al. [35] and Doolani et al. [34] introduce the intriguing dimension of user satisfaction, engagement, and learning. These studies, along with Peruzzini et al. [20], underscore the importance of assessing user experience and the effectiveness of training methods. By measuring factors like enjoyment, comfort, and readiness level, they aim to optimize training processes, enhance learning outcomes, and ensure user satisfaction.

Hoesterey and Onnasch, and Kuts et al. [29,30] dive into safety assessment and user health in high-risk environments, mirroring the safety considerations in Havard et al. [45]. They collectively strive to identify and manage risks, optimize safety, and improve decision making in challenging scenarios.

As it can be seen, the approaches to these measurements vary significantly. Although all this information can be found in Appendix A (Table A2), Table 5 shows a part of the whole table to get an idea of the data collected.

Table 5. Identified human factors, metrics, and data collection methods (the full table can be found in Appendix A (Table A2); n/d = not defined).

Factor	Hedonic vs. Pragmatic	Categories	Metrics	Unit of Measurement	Technique/Data Collection Method	Paper
Acceptability	Hedonic	Cognitive	Suitability	Interview	Interview	[27]
Attention	Hedonic	Physiological	Gaze	Time per object	Eye tracking data analysis	[30]
Comfort	Hedonic	Physiological	Brain Activity (Alpha)	n/d	EEG	
Comfort	Pragmatic	Process	Spatial properties	Self-generated questionnaire (1–5 and 1–10)	Questionnaire	
Comfort	Pragmatic	Process	Aesthetic properties	Self-generated questionnaire (1–5 and 1–10)	Questionnaire	[33]
Comfort	Hedonic	Cognitive	Likeability	Self-generated questionnaire (1–5 and 1–10)	Questionnaire	
Effectiveness	Pragmatic	Cognitive	Need of support	Number of times asked for help	User observation	
Effectiveness	Pragmatic	Process	Workarounds created	Number	User observation	
Effectiveness	Pragmatic	Physiological	Gaze	Number	Eye tracking data analysis	[20]
Effectiveness	Pragmatic	Physiological	Heat map (dimension of the area with visual interaction	Area (mm2)	Eye tracking data analysis	
Effectiveness	Pragmatic	Process	Average training time	Time	Manually–stopwatch	
Effectiveness	Pragmatic	Process	Average tutorial time	Time	Manually–stopwatch	[44]
Effectiveness	Pragmatic	Process	Average assessment time	Time	Manually–stopwatch	
Effectiveness	Pragmatic	Process	Effectiveness	Subjective judge (1–5 point scale)	Questionnaire	
Efficiency	Pragmatic	Process	Task execution time	Time	Digital simulation analysis	
Efficiency	Pragmatic	Physiological	Postural comfort	Comfort level (1–7, 1–4, 1–11) according to different methods (RULA, OWAS, REBA…)	Digital simulation analysis	[20]

4. Discussion

The present scoping review aimed to transform the broad and diverse literature of the assessment of human factors in industrial VR environments. Through the analysis of these 24 selected papers, a critical review of the various key issues is followed by a detailed discussion on the main topics.

4.1. Lack of Consistency About the Terms

One of the great discoveries in this review has been, in addition to seeing the many ways that authors use to measure human factors, the lack of consistency in the taxonomy of terms. Terms have been found in some studies to be classified as human factors but appearing as metrics in others.

As an example, the term comfort appears in the literature as a human factor [33], but it also appears as a metric to measure ergonomics [44]. On the other hand, learnability can be found as an independent factor [23,24,27,34,36,44], but it can also be found as a metric of usability [27,34,35,46]

Another more complex case is effort and performance factors. These terms appear as human factors in Bernard et al. [37] and [22,26], respectively. Nevertheless, both effort and performance appear directly dependent on workload, since in the NASA TLX tool

they appear as such [21,22,26,28,31,37]. Effort, performance, and workload also appear dependent on stress level in the literature [30].

Efficiency also appears as a human factor, [20,37,44,46] but also dependent on UX [24]. As can be seen, even the interdependencies between factors and metrics are not clear. Sometimes this is due to the tools used that already define how to classify these elements. These mentioned relationships can be seen in Table 6.

Table 6. Inconsistencies between term relationships.

Human Factor	Metric	Author
Comfort	Brain activity (Alpha) ...	[33]
Ergonomics	Comfort	[44]
Learnability	Ease of use ...	[24,34]
Usability	Learnability ...	[27,34,35,46]
Stress	Effort Performance Workload ...	[26]
Workload	Effort Performance ...	[29,32,38]
Effort	Perceived physical exertion ...	[45]
Performance	Errors	[29,32]
Efficiency	Assessment scores ...	[20,37,44,46]
User Experience	Efficiency ...	[30]

Reviewing the literature, a lack of consistency is also present in the collected data on the factors most frequently used in the experiments. They are shown in Table 7, together with the number of metrics and the metrics themselves classified. Usability, workload, ergonomics, learnability, and user experience are the most repeated. It is worth noting the 32 related metrics classified as cognitive and pragmatic in usability. This is because most of the papers in there have used the SUS scale [27,34,35,46], which has seven metrics of this type, but regarding the overall number of papers that include these human factors, these are the most frequent ones.

It can also be observed that hedonic metrics in general are considerably lower than pragmatic metrics in both the process and physiological categories. Regarding cognitive metrics, except for the usability factor that is related to pragmatism, the rest is dominated by hedonic metrics. It can also be observed how in this field ergonomics tends to be measured pragmatically with metrics of physiological characteristics.

The observed variability in the classification and measurement of human factors within the topic of XR underscores the field's low maturity and the complexity inherent in capturing the multifaceted nature of human experiences. The lack of standardization in

terms may stem from the interdisciplinary nature of XR research, where fields like cognitive psychology, ergonomics, human–computer interaction, and industrial design converge. Each discipline brings its own lexicon and methodological approaches, contributing to the richness of perspectives but also to terminological discrepancies.

Table 7. Most frequent human factor metrics classified.

Human Factor	Type	Process	Physiological	Cognitive	Other
Usability	Pragmatic	14	1	32	-
	Hedonic	6	-	5	-
Workload	Pragmatic	18	15	6	-
	Hedonic	-		9	-
Ergonomics	Pragmatic	-	10	-	-
	Hedonic	-	2	-	-
Learnability	Pragmatic	7	-	2	-
	Hedonic	-	-	8	1
User Experience	Pragmatic	3	1	3	-
	Hedonic	-	-	4	-

The necessity for standardization is a double-edged sword. On the one hand, a unified taxonomy would facilitate clearer communication among researchers, practitioners, and stakeholders, enabling more effective collaboration and cumulative knowledge building. On the other hand, the dynamism of the field may require a certain level of flexibility to accommodate new discoveries and technological advancements.

Looking ahead, it is plausible that the field will trend towards a consensus on certain core terms, driven by a growing body of interdisciplinary research and the establishment of best practices. However, complete standardization may remain elusive, reflecting the evolving and adaptive nature of human factors research in XR. As the technology matures and becomes more embedded in industrial applications, the demand for standardized metrics that can reliably predict user outcomes will likely increase, potentially leading to more rigorous methodological frameworks and measurement tools.

In this scenario, the future of XR research could involve a harmonious balance between standardized methodologies and the flexibility to explore the unique nuances of human interaction with emerging technologies. This balance would aim to preserve the innovative spirit of the field while providing the structure needed for its growth and integration into industrial and other practical applications.

4.2. Correlation Among Factors

Regarding industrial VR experiences, researchers have conducted studies to uncover correlations among various human factors, showing the intricate dynamics of user interactions, subjective responses, and physiological indicators within these immersive environments. This comprehensive analysis synthesizes these correlations into a cohesive understanding of how users navigate and respond to the challenges presented in VR-based industrial scenarios.

One critical correlation observed pertains to fixation metrics and hazard perception. Ref. [31] noted that fixation frequency and fixation duration increased as participants encountered more hazards in their workplace. This increase in fixation metrics suggests a heightened mental workload, as users focus more intently on potential risks and challenges in their virtual surroundings. Furthermore, the study found that saccade duration decreased while saccade amplitude increased with exposure to workplace hazards. While the

shorter saccade duration aligns with earlier research [47] the increase in saccade amplitude represents a novel finding in this context.

Subjective workload assessment also plays a crucial role in understanding how users perceive hazardous scenarios. In Das et al. [31], participants reported increased mental and temporal demands as workplace hazards escalated. However, physical demand, effort, and frustration remained largely unaffected by the hazardous scenarios, highlighting the nuanced nature of subjective responses in VR.

Building upon the relationship between subjective responses and activity level, the study observed that hazardous scenarios and activity levels showed a consistent correlation with subjective workload. Interestingly, the trial's impact on subjective responses appeared to be insensitive. Furthermore, the NASA-TLX score, a measure of total workload, exhibited significant correlations with various levels of mental workload. Notably, the highest total workload was observed during mixed experiment combinations involving highly hazardous scenarios, complex activity levels, and the initial trial.

The study reported strong and significant correlations between NASA-TLX scores and eye movement metrics [31]. Saccade amplitude displayed a weak and non-significant relationship with performance measures, while frustration did not correlate with any eye movement metrics or performance measures. Subjective workload evaluation appears to be influenced by a user's sense of success or self-conception of their performance, underlining the strong connection between these feelings and actual task performance.

Transitioning to the performance and user experience of VR interfaces, Kuts et al. [30] found that VR interfaces outperformed traditional methods in terms of time spent placing objects, with average times reduced after users gained familiarity with the virtual environment. However, the VR scenario induced higher levels of anxiety and greater mental and physical demands on operators. Remarkably, these factors did not significantly affect physiological stress levels, indicating a complex relationship between perceived stress and physiological responses.

The introduction of digital twin (DT) systems in VR also sparked interest in user acceptance and task execution performance. Kuts et al. [30] reported promising results, supporting the notion that VR can be a valuable alternative to traditional robot programming interfaces, thanks to its acceptability and overall task execution performance.

Moreover, eye tracking studies conducted by Kuts et al. [30] revealed that users tended to direct their attention more frequently towards the main robot VR user interface than the robot twin. This skewed attention distribution may be attributed to the perceived safety of the main interface, emphasizing the critical role of user perception in influencing attention and focus within VR environments. Intriguingly, attention-tracking results suggested that the type of interface interaction in VR could further influence user attention and focus, highlighting the multifaceted nature of user engagement in immersive virtual environments.

4.3. Overlooking of Technological and Task Skills

In VR testing environments, overlooking the technological and task skills of participants can significantly skew the results of usability studies. This oversight is particularly problematic due to, among other factors, the digital divide, which encompasses disparities in access to and familiarity with technology [48]. Participants with limited technological experience might face initial difficulties that more experienced users do not, leading to an unequal assessment of the VR system's usability. This disparity can mask genuine usability issues and mislead researchers about the true efficacy of the VR system. In this sense, different strategies have been identified to minimize the impact.

The variability in participant skill levels aggravates these issues. Participants come with diverse backgrounds and varying degrees of familiarity with technology, which can substantially influence the outcomes of usability studies. This happens in the study by Das et al. [31], which highlights the need to account for diverse user backgrounds to obtain accurate usability assessments. Novice users, those with limited experience, often face a steep learning curve. They may struggle with basic functionalities initially but show substantial improvement as they become more familiar with the system. This learning trajectory needs to be carefully documented and understood to avoid misinterpreting improvements in usability solely as a result of iterative design changes rather than increased user familiarity.

Conversely, experienced users might demonstrate proficiency from the outset, but their feedback could highlight more advanced issues that novice users might not encounter. This dichotomy between novice and experienced users can lead to a diverse range of feedback that needs to be carefully analyzed. Novice users might report basic usability issues, while experienced users might focus on more nuanced aspects of the interface, such as efficiency and advanced functionalities. This was observed in the study by Nenna et al. [22], where task familiarity significantly impacted user feedback.

Iterative testing, a common practice in prototype development, further complicates the situation. This process involves testing a prototype, gathering feedback, making improvements, and then retesting. While essential for refining and improving a product, iterative testing also means that participants become increasingly familiar with the task and the prototype. This familiarity can influence their feedback and performance, potentially leading to skewed results. For example, when users test a prototype multiple times, their feedback may become more positive simply because they are more comfortable and adept at using the prototype. This can create a false sense of improvement, as the enhancements in user performance and satisfaction may be more attributable to increased familiarity rather than genuine improvements in the prototype's usability. Research on cognitive testing has shown that significant practice effects can emerge over time due to repeated testing, emphasizing the need for careful analysis of repeated testing data to distinguish between genuine usability improvements and mere familiarity effects [49].

These dynamics have significant implications for the design of user testing protocols. Researchers must carefully consider whether to allow for prior familiarization with the task or technology. One approach is to conduct initial baseline testing with participants who have no prior exposure to the task, followed by additional testing sessions to observe changes over time. This approach can help isolate the effects of familiarity and provide a clearer picture of genuine usability improvements. Additionally, when mixing participants with different skill levels, it is essential to segment the data accordingly. By analyzing the performance of novice and experienced users separately, researchers can gain a more nuanced understanding of how different user groups interact with the technology. This segmentation helps in identifying specific areas where novice users struggle and where experienced users find the interface lacking in advanced functionalities. For instance, in the study by Bu et al. [32], segmenting feedback from different skill levels provided richer insights into user interactions and highlighted specific areas for improvement. Similarly, Rogers et al. [36] emphasized the importance of evaluating the effectiveness of VR as a learning tool among users with different levels of experience, further illustrating the varied impacts of user expertise on usability assessments.

In contrast, several studies do not take user expertise into account, which can lead to incomplete or skewed findings. For example, the study by Havard et al. [45] on digital twin and virtual reality co-simulation does not consider the participants' expertise levels, potentially overlooking how varying familiarity with VR technology can affect usability

outcomes. Similarly, the study by Morosi et al. [26] on configuring a VR simulator for ergonomic evaluation also lacks consideration of user expertise, which might result in an underestimation of the learning curve effects on user performance and satisfaction.

Another approach is to use a combination of qualitative and quantitative methods to capture a comprehensive view of user experiences [50]. Qualitative feedback from interviews and observations can provide context to quantitative data, helping to explain why certain patterns emerge in the user testing results. For example, novice users might express frustration in interviews, which can explain high error rates observed in quantitative metrics. Similarly, experienced users might suggest advanced features during interviews that are not apparent in their quantitative performance data. Furthermore, researchers can design testing protocols that account for the learning curve. This can involve conducting baseline tests to establish initial performance levels and then monitoring changes over subsequent sessions. By comparing these results, researchers can better isolate the effects of familiarity from genuine usability improvements. This approach was highlighted in research on cognitive testing, where significant practice effects were observed over time, emphasizing the need for careful analysis of repeated testing data [49]. Additionally, Hoesterey et al. [29] discussed how manipulating situational risk in experimental paradigms can help in understanding the impacts of user expertise on task performance in VR environments.

In conclusion, the variability in participant skill levels is a critical factor in user testing that requires careful consideration. By designing thoughtful testing protocols and segmenting participants, researchers can better manage the influence of prior familiarity and skill levels on user testing outcomes. More extensive research in this area will continue to enhance our understanding and enable the development of best practices for conducting effective and reliable user tests.

4.4. Integration of Multimodal Data in Human Factors Analysis

The assessment of human factors in VR environments for Industry 5.0 leverages various types of data to provide a comprehensive understanding of user interactions and experiences. These data types can be broadly categorized into physiological, cognitive, and performance metrics as previously showed in Table 4.

Integrating these diverse data sources can lead to a more holistic understanding of human factors in VR environments. By combining physiological data with cognitive assessments and performance metrics, researchers can gain richer insights into how users interact with VR systems [22]. Comparing user-reported cognitive load with physiological data, like heart rate variability and skin conductance, can help confirm and strengthen the study's findings. If a user reports high mental workload and this is corroborated by an elevated heart rate and EDA, the assessment is more robust [28]. Multimodal data integration allows for the creation of detailed user profiles that include cognitive states, emotional responses, and physical performance. This can inform personalized VR training programs tailored to individual needs and capabilities. Understanding how physiological stress correlates with specific VR tasks can guide the design of more ergonomic and user-friendly interfaces. For instance, if high stress levels are associated with certain interaction patterns, designers can modify the interface to mitigate these issues [26]

Despite the benefits, integrating multimodal data presents several challenges. Aligning data from different sources, each with its own temporal resolution and format, is technically complex. Solutions include using synchronized logging systems and developing algorithms to align data streams based on common timestamps [27]. Interpreting data from different modalities requires interdisciplinary expertise. Collaboration between experts in physiology, psychology, and human–computer interaction is essential to make sense of the integrated data. The sheer volume of data can be overwhelming. Advanced data analytics

and machine learning techniques can help manage and analyze large datasets to extract meaningful patterns.

While advanced technologies enhance data collection, they can also intrude on the user experience. Devices like EEG caps, motion capture suits, and multiple sensors can be cumbersome and affect natural interactions within the VR environment. Wearing multiple sensors and heavy equipment can cause physical discomfort, potentially influencing the user's performance and subjective experience [51] Lightweight, less intrusive alternatives are being developed, such as wearable sensors integrated into clothing or non-contact biometric sensors. The awareness of being monitored can alter user behavior, known as the Hawthorne effect [52]. Ensuring user comfort and reducing the visibility of monitoring devices can help mitigate this issue. Current VR systems are evolving, but the need for robust data collection often requires trade-offs with user comfort and naturalistic interaction. Future advancements should focus on minimizing intrusiveness while maintaining data accuracy, such as integrating biometric sensors directly into VR headsets.

5. Research Gaps and Future Research Directions

In the context of evaluating industrial VR experiences and the inclusion of human factors, there are several research gaps and future research directions that need to be addressed.

5.1. Lack of Standardized Method for Measuring Human Factors

There is a lack of standardization of how to measure human factors according to the needs of the experiment, and this can lead to interpreting terms in a different way, using different scales, or even using methods that are not suitable for these technologies. Sometimes there is a tendency to use standardized tools that have been applied in other fields beyond VR. This can be a mistake if the tool is not validated for this technology. This has happened before with the SSQ (Simulation Sickness Questionnaire) [53], which was initially developed for aviation pilots and their simulators, which made the defined zero baseline very high for the rest of the simulators. There are already papers where it is used, but some authors [53,54] have sought alternatives or adaptations for applications, such as the VRSQ [55]. The same applies to the NASA-TLX and its increasingly frequent adaptations such as the SIM-TLX [56]. There are also several factors that affect this lack of standardization that should be considered for future research.

Adapting tools from other contexts to VR may introduce inaccuracies. Future research endeavors should prioritize the development and validation of standardized measurement tools tailored explicitly to industrial VR, accounting for its unique characteristics for the precise assessments of human factors.

Future research should focus on establishing general guidelines and a standardized taxonomy that categorizes human factor metrics (e.g., cognitive, physiological, performance, and subjective measures). This framework should encourage a balance between objective and subjective methods while validating tools specifically for VR contexts, rather than adapting measures from other fields without proper assessment. Collaborative efforts among researchers and practitioners can help define clear benchmarks and shared repositories of validated tools, ensuring that assessments align with the unique characteristics and demands of VR environments.

5.2. Sampling Bias

Several studies underscore limitations related to the sample size and bias of their findings. Das et al. [31] acknowledge a relatively small sample size, while Nenna et al. [22] note that their conclusions are drawn from a limited pool of young users with limited experience, emphasizing the need for wider samples. Havard et al. [45] similarly face the challenge of a modest number of users in their study. The authors of Ref. [24] discuss the limitations of a small sample size and the inability to claim representativeness. This gives rise to several problems, among them that the studies have a biased sample and sometimes distorted results, as there are uncontrollable variables in both the system and the individuals. As mentioned above, it is worth noting the difference between studies in terms of sample size, skills, and the gender of the participants.

Dealing with sampling bias is critical in industrial VR studies, as it can lead to skewed results and limited generalizability. Commonly observed small sample sizes may not adequately represent the diverse range of potential users in industrial settings. Limited demographic diversity in study participants can also introduce bias. Future research should prioritize larger and more diverse samples, encompassing participants of different ages, genders, and backgrounds, to ensure findings are representative and applicable to the broader industrial workforce.

5.3. Technology and Hardware Constraints

Limitations related to technology and hardware are common concerns in the VR research landscape. Variations in proprietary algorithms used for eye-tracking data acquisition are also highlighted [20,22]. Bernal et al., and Peruzzini et al. [20,35] point out complexities in system setup, integration issues, and the challenge of hardware obsolescence. Dado et al. [57] mentions the use of software not specifically designed for training, which can impact the effectiveness of VR experiences. Effective implementation of haptic feedback also emerged as a challenge in several studies. Morosi and Caruso [26] discuss the difficulties in achieving meaningful haptic feedback, emphasizing that different solutions may lead to the same result, but their selection is not trivial.

It also happens that more and more devices are emerging with new features, such as Apple's recent announcement of the Apple Vision Pro [58], which already proposes a hybridization between augmented and mixed reality. But this development is not only happening with the approach of the technology itself, but with the features that allow for the collection of more information from the user and in a not so invasive way. Up to now, the testing itself has been conditioned by the intrusiveness of the measurement elements as mentioned in Section 4.4. EEGs, IMUs, and other devices are extra elements that must be added to the user, and this can hinder the execution of activities, hindering movements, decreasing attention, etc. Different authors mention their concern about this issue and seek a balance between optimal data collection and the level of intrusion of the person. For example, the use of an invasive eye tracker which was very light and comfortable to wear in some of the experiments [31]. Fortunately, this is changing, and as VR devices continue to evolve, emerging technologies offer promising solutions to address many of the challenges discussed.

For instance, generative AI can enhance data collection and analysis by adapting environments in real time, improving user experiences without requiring additional hardware. Similarly, advancements in haptic technology aim to achieve more precise and realistic tactile feedback, addressing the difficulties highlighted in system integration and meaningful feedback design. Moreover, the increasing ability of non-invasive sensors embedded in HMDs to track body movements, facial expressions, and voice data can reduce the reliance on intrusive measurement tools like EEGs and IMUs. These developments not only minimize user discomfort but also streamline data acquisition, allowing researchers to balance accuracy and usability. Combined, these innovations mark a shift toward more seamless, robust, and user-centered methods for evaluating VR systems.

5.4. VR Environment Tailoring

The imperative to tailor VR environments to the tasks is a concept gaining traction, with the specificity of such customization being a decisive factor in the utility and effectiveness of VR applications in industrial settings. Morosi and Caruso [26] have highlighted the criticality of configuring VR experiences to support the exact nature of the tasks which operators are required to perform. This customization ensures that the technological capabilities of the VR system align closely with the cognitive and physical demands of these tasks, enhancing both performance and user satisfaction.

Zhang et al. [46], further reinforce this notion by acknowledging the limitations of their study, which was constrained to a single task and object. This admission points to a broader recognition that the scope and transferability of findings in VR research are significantly influenced by the extent to which the environment and tasks have been tailored. Such specificity in VR environment tailoring is a practical application of the technology task fit (TTF) model [59] which suggests that the effectiveness of technology is maximized when it is closely aligned with the demands of the task.

For VR technologies to be most effective, they must be adaptable to a variety of tasks and user scenarios. The development of VR systems should thus be informed by a thorough analysis of the intended tasks, incorporating adaptability and customization into the design process. A VR environment that is too rigid or narrowly focused may not offer the flexibility required to accommodate diverse or evolving industrial tasks.

This implies that future research and development in VR should not only focus on technological advancement but also on a deeper understanding of the tasks themselves. This includes recognizing the complexity, interdependencies, and nuances of tasks to ensure that VR systems provide relevant functionalities and interfaces that enhance task performance. This approach would ensure that VR technologies are not only advanced in their capabilities but are also relevant and effective tools that support users in achieving their goals within the industrial context.

6. Limitations and Conclusions

In the context of this review, several limitations warrant careful consideration, such as those which follow:

The systematic literature review (SLR) methodology, esteemed for its rigor, does not guarantee the exhaustive identification of all publications within a specific research domain, as evidenced by prior work. This inherent incompleteness is an acknowledged aspect of the SLR process.

The deliberate decision to exclusively focus on peer-reviewed articles introduces a notable limitation. By applying this, the potential exclusion of valuable case studies presented at conferences arises, compromising the comprehensiveness of the study given that conference case studies constitute a distinct and valuable source of insights.

Despite conscientious efforts to maintain objectivity throughout the review process, the possibility of introducing bias in certain instances is recognized as a challenge. This is an inherent aspect of research endeavors of this nature, where subjective judgments and interpretations may influence overall findings. The selection of databases, though thoughtfully undertaken to encompass a broad spectrum of the research area, is not without its limitations. While the inclusive search strategy aimed to capture a comprehensive range of relevant literature, the possibility remains that additional databases could have revealed more significant articles for inclusion, thereby enhancing the depth of the study.

In establishing quality assessment (QA) criteria, the choice of QA questions and the setting of a QA threshold at three were strategic decisions. It is crucial to acknowledge that alternative QA questions or a lower threshold could have yielded substantially different outcomes in the SLR. However, the chosen QA criteria, along with the specified threshold, contributed to the identification of a collection of high-quality papers, a determination substantiated by the literature characterization.

A final limitation pertains to the decision to restrict results exclusively to publications in English and Spanish. While this choice streamlines the review process, it introduces a language restriction that may exclude valuable contributions in other languages. This linguistic limitation poses a potential barrier to a more inclusive and diverse representation of perspectives within the study.

In recognizing and addressing these limitations, we establish a clear understanding of the study's scope and constraints. This critical evaluation sets the stage for our conclusions, where we synthesize the findings of the systematic literature review, offering valuable insights into the assessment of human factors in industrial VR environments and identifying opportunities for future research within the evolving landscape of Industry 5.0.

A key observation in this review underscores the lack of consistency in the taxonomy of terms used to measure human factors. The terminology considered as human factors in some studies is utilized as a metric in others. Additionally, the interdependencies between factors and metrics are often unclear, partly due to the varied tools employed to classify these elements.

The review further highlighted the prevalence of certain factors in experimental studies, with usability, workload, ergonomics, learnability, and user experience emerging as the most frequently addressed factors. Usability featured the most associated metrics, especially those of cognitive and pragmatic nature. Notably, hedonic metrics were considerably less prominent than pragmatic metrics, particularly in process and physiological categories. Ergonomics tended to be assessed pragmatically, focusing on metrics related to physiological characteristics. In the field of industrial VR experiences, researchers have delved into uncovering correlations among diverse human factors. These studies illuminate the intricate dynamics of user interactions, subjective responses, and physiological indicators within immersive environments.

Although the rapid evolution of VR technology brings opportunities, it also brings some challenges. The swift obsolescence of VR technology may result in compatibility issues with older hardware and software used in research studies. Challenges persist in achieving meaningful haptic feedback, and the intrusiveness of some data collection methods, such as external devices like EEGs and IMUs, remains a concern. Researchers should vigilantly assess and adapt to evolving VR technology and hardware, aiming to minimize intrusiveness and enhance haptic feedback for a more realistic industrial VR experience.

The findings of this SLR, set against the backdrop of the imminent Industry 5.0 era, highlight significant challenges and opportunities in industrial VR environments. Key among these is the inconsistent use of terminology and metrics in measuring human factors, underscoring the need for standardization in this rapidly evolving field. This aligns with Industry 5.0's focus on integrating advanced technologies like VR while emphasizing human-centric approaches, leading to more efficient, inclusive, and human-centered industrial environments.

Appendix A

In this section, the three abovementioned tables are presented in full. Table A1 lists the different devices identified in the literature, categorized accordingly. Table A2, on the other hand, compiles all identified human factors, along with their respective metrics, measurement units, and classified data collection methods.

Table A1. Devices used in the literature.

Type	Main Device	Extra Elements	Category	Ref.
CAVE	CAVE	Wireless EEG device	Biosensor	[33]
		Microsoft Kinect	Motion capture	
		Infrared cameras	Motion capture	[20]
		BioHarness 3.0	Biosensor	
		Tobii Pro Eyeglasses 2	Biosensor	
		Microsoft Kinect	Motion capture	[27]
		ART Tracking	Motion capture	
		Motion Capture System	Motion capture	[35]
		Joystick	Controller	

Table A1. *Cont*.

Type	Main Device	Extra Elements	Category	Ref.
HMD (Head mounted display)	HTC Vive [61]	Headphones	Sound	[34]
		n/d (not defined)	n/d	[23]
		Microsoft Kinect	Motion capture	[57]
		n/d	n/d	[21]
	HTC Vive Pro [62]	3D limb-sensing device with visual, tactile and auditory simulation	Physical/Haptic	[32]
		Polar watch	Biosensor	
	HTC Vive Pro Eye [63]	n/d	n/d	[22]
		Headphones	Sound	[29]
		Heart rate monitor	Biosensor	[30]
		Empatica E4	Biosensor	[28]
		HTC Vive Trackers 3.0	Motion capture	
		BioHarness 3.0	Biosensor	
		Leap motion	Motion capture	[19]
		X Sens	Motion capture	
	n/d	Percepcion Neuron Pro suit	Motion capture	[45]
	Oculus Quest 2 [64]	Robotic arm	Physical/Haptic	[46]
	Oculus Rift [65]	Tobii Pro eyeglasses II	Biosensor	[31]
		Keyboard	Controller	[36]
		Leap motion	Motion capture	[26]
		5.1 surround	Sound	
		Haptic master	Physical/Haptic	
		Percepcion Neuron Lite	Motion capture	[44]
		Eyetracker	Biosensor	[41]
		Microsoft Kinect	Motion capture	
		Leap motion	Motion capture	[25]
		n/d	n/d	[24]
		PMU (physical mockup)	Physical/Haptic	[37]

Table A2. Human Factors, metrics classification and data collection methods.

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Acceptability	Suitability	Hedonic	Cognitive	Interview	Interview	[27]
Attention	Gaze	Hedonic	Physiological	Time per object	Eye tracking data analysis	[30]
Comfort	Brain Activity (Alpha)	Hedonic	Physiological	ND (not defined)	EEG	[33]
	Spatial properties	Pragmatic	Process	Self-generated questionnaire (1–5 and 1–10)	Questionnaire	[33]
	Aesthetic properties	Pragmatic	Process	Self-generated questionnaire (1–5 and 1–10)	Questionnaire	[33]
	Likeability	Hedonic	Cognitive	Self-generated questionnaire (1–5 and 1–10)	Questionnaire	[33]

Table A2. *Cont.*

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Effectiveness	Need of support	Pragmatic	Cognitive	Number of times asked for help	User observation	[20]
	Workarounds created	Pragmatic	Process	Number	User observation	[20]
	Gaze	Pragmatic	Physiological	Number	Eye tracking data analysis	[20]
	Heat map (dimension of the area with visual interaction	Pragmatic	Physiological	Area (mm^2)	Eye tracking data analysis	[20]
	Average training time	Pragmatic	Process	Time	Manually stopwatch	[44]
	Average tutorial time	Pragmatic	Process	Time	Manually stopwatch	[44]
	Average assessment time	Pragmatic	Process	Time	Manually stopwatch	[44]
	Effectiveness	Pragmatic	Process	Subjective judge (1–5 point scale)	Questionnaire	[41]
Efficiency	Task execution time	Pragmatic	Process	Time	Digital simulation analysis	[20]
	Postural comfort	Pragmatic	Physiological	Comfort level (1–7, 1–4, 1–11) according to different methods (RULA, OWAS, REBA…)	Digital simulation analysis	[20]
	Visibility	Pragmatic	Physiological	Heuristic evaluation of field of view (1–10)	Digital simulation analysis	[20]
	Task execution time	Pragmatic	Process	Time	Manually stopwatch	[37]
	Assessment scores	Pragmatic	Process	Score	Manually stopwatch	[44]
	Task execution time	Pragmatic	Process	Time	Digital simulation analysis	[20]
Effort	Perceived physical exertion	Pragmatic	Physiological	Borg RPE (0–10 point scale)	Questionnaire	[37]
Ergonomics	Body position	Pragmatic	Physiological	RULA score	Position tracking	[45]
	Posture	Pragmatic	Physiological	Time	Reach envelope (analysis option)	[37]
	Posture	Pragmatic	Physiological	RULA score	Questionnaire	[37]
	Musculoskeletal symptoms	Pragmatic	Physiological	Nordic questionnaire (1–7 point scale)	Questionnaire	[37]
	Postural overload	Pragmatic	Physiological	RULA score	Questionnaire	[28]
	Body part discomfort	Pragmatic	Physiological	Body part discomfort scale (1–5 point scale)	Questionnaire	[28]
	Comfort	Hedonic	Physiological	Subjective judge (1–5 point scale)	Questionnaire	[44]
	Posture	Pragmatic	Physiological	OWAS (Ovako Working posture Analyzing System)	Worksheet	[19]
	Comfort	Hedonic	Physiological	REBA (Rapid Entire Body Assessment)	Worksheet	[19]
	Physical workload	Pragmatic	Physiological	EAWS (European Assembly Work-Sheet)	Worksheet	[19]
	Human performance (Joint angles)	Pragmatic	Physiological	DHM (Change in posture)	Motion capture	[21]
	Vision performance	Pragmatic	Physiological	Obscuration zone analysis (% area of vision blocked)	Computing	[21]

Table A2. *Cont.*

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Game Experience	Competence	Pragmatic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Sensory and imaginative immersion	Pragmatic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Flow	Pragmatic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Tension-Annoyance	Pragmatic	Physiological	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Challenge	Pragmatic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Negative affect	Hedonic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Positive affect	Hedonic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Positive Experience	Hedonic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Negative Experience	Hedonic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Tiredness	Pragmatic	Physiological	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
	Returning to reality	Pragmatic	Cognitive	GEQ (Game Engagement Questionnaire 0–4)	Questionnaire	[35]
Immersion	Immersion	Pragmatic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[44]
	Immersion	Pragmatic	Cognitive	Self generated questionnaire (0–7)	Questionnaire	[27]
Learnability	Task execution accuracy	Pragmatic	Process	Score (anova comparison)	Assessment/Exam	[36]
	Task execution accuracy	Pragmatic	Process	mm	Comparative: optimal vs. result	[34]
	Task execution time	Pragmatic	Process	Time	Manually-stopwatch	[34]
	Difficulties in understanding or execution	Pragmatic	Process	Number of times asked for help	User observation (video recordings)	[24]
	Ease of use	Pragmatic	Process	Likert scale (1–7)	Questionnaire	[24]
	Errors	Pragmatic	Process	Number	User observation (video recordings)	[24]
	Intention of use	Hedonic	Cognitive	Likert scale (1–7)	Questionnaire	[24]
	Motivation in the learning process	Hedonic	Cognitive	Likert scale (1–7)	Questionnaire	[24]
	Other anomalies	Other	Other	ND	User observation (video recordings)	[24]

Table A2. *Cont.*

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Learnability	Subjective assessment of the learning success	Hedonic	Cognitive	Likert scale (1–7)	Questionnaire	[24]
	Task execution time	Pragmatic	Process	Time	User observation (video recordings)	[24]
	Technology acceptance	Hedonic	Cognitive	TAM (Technology Acceptance Model)	Questionnaire	[24]
	Usefulness	Hedonic	Cognitive	Likert scale (1–7)	Questionnaire	[24]
	Potential for VR training development	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[44]
	Learnability	Pragmatic	Cognitive	QUIS (Questionnaire for User Interface Satisfaction) (0–9)	Questionnaire	[27]
Learnability/ Conceptual understanding	Learnability	Pragmatic	Cognitive	Knowledge of learning objectives	Assessment/Exam	[23]
Learnability/ Self efficacy	Ability to understand and explain the concepts	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[23]
	Familiarity with and comfort in applying the concepts	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[23]
Memorability	Task execution accuracy	Pragmatic	Process	mm	Comparative: optimal vs. result	[34]
	Task execution time	Pragmatic	Process	Time	Manually stopwatch	[34]
	Number of repetitions of the training	Pragmatic	Process	Number	Counter	[34]
	Task execution time	Pragmatic	Process	Time	Manually stopwatch	[34]
Mental workload	Electrodermal activity	Pragmatic	Physiological	ND	EDA Monitor	[28]
	Heart rate	Pragmatic	Physiological	Time	Heart Rate Monitor thoracic band	[28]
	Pupillometry	Pragmatic	Physiological	Pupil size variation/time	Eye tracking data analysis	[28]
Performance	Number of Errors	Pragmatic	Process	Percentage of wrong answers	User observation	[22]
	Task execution time	Pragmatic	Process	Time	Software/video	[22]
	Errors	Pragmatic	Process	Number	3D position and trajectory	[26]
	Task execution time	Pragmatic	Process	Time	Manually stopwatch	[26]
	Task execution time	Pragmatic	Process	Time	Software/video	[30]
Presence	Realism	Hedonic	Process	Witmer and Singer presence questionnaire adaptation scale (1–7)	Questionnaire	[46]
	Possibility to act	Pragmatic	Process	Witmer and Singer presence questionnaire adaptation scale (1–10)	Questionnaire	[46]
	Self-evaluation of performance	Pragmatic	Process	Witmer and Singer presence questionnaire adaptation scale (1–10)	Questionnaire	[46]
	Haptic	Pragmatic	Physiological	Witmer and Singer presence questionnaire adaptation scale (1–10)	Questionnaire	[46]

<div align="center">

Table A2. *Cont.*

</div>

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Risk detection	Number of hazards identified	Pragmatic	Process	Number	Assessment/Exam	[57]
	Number of correct hazards identified	Pragmatic	Process	Number	Assessment/Exam	[57]
Safety	Workplace safety	Pragmatic	Process	APACT check list (0–10 point scale)	Checklist	[37]
Satisfaction	Subjective visibility (can you properly see everything you need?)	Pragmatic	Physiological	Subjective judge (1–5 point scale)	Questionnaire	[29]
	Accessibility (can you properly reach everything you need?)	Pragmatic	Physiological	Subjective judge (1–5 point scale)	Questionnaire	[29]
	Mental demand	Pragmatic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[29]
	Emotional (Is the stress to accomplish the task appropriate?)	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[29]
	Perceived comfort (Are you feeling in a comfortable position?)	Hedonic	Physiological	Subjective judge (1–5 point scale)	Questionnaire	[20]
Situational Risk	Anxiety	Hedonic	Cognitive	State-Trait_anxiety Inventory (STAI-S) (8 point Likert scale)	Questionnaire	[29]
	Hesitation time	Pragmatic	Process	Time	Manually stopwatch	[29]
	Caution time (time spent outside of the expected zone)	Pragmatic	Process	Time	Manually stopwatch	[29]
	Valence	Hedonic	Cognitive	SAM (Self Assessment Manikin) (9 point scale)	Questionnaire	[29]
	Arousal	Hedonic	Cognitive	SAM (Self Assessment Manikin) (9 point scale)	Questionnaire	[29]
	Difficulty stepping out	Pragmatic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[29]
	Avoidance falling off	Pragmatic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[29]
	Dared to step off	Hedonic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[29]
	Feel risk	Hedonic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[29]

<div align="center">

Table A2. *Cont.*

</div>

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Stress level	Mental demand	Pragmatic	Cognitive	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Physical demand	Pragmatic	Physiological	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Temporal demand	Pragmatic	Process	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Performance	Pragmatic	Process	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Effort	Pragmatic	Physiological	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Frustration	Hedonic	Cognitive	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Total workload	Pragmatic	Process	NASA-TLX Scale (1–10)	Questionnaire	[26]
	Heart rate	Pragmatic	Physiological	Bpm	Heart rate monitor	[26]
	Electrodermal activity	Pragmatic	Physiological	ND	EDA Monitor	[28]
	Inter-beat intervals (heart)	Pragmatic	Physiological	Time	Heart rate monitor thoracic band	[28]
Suitability and Relevance of use	Suitability and relevance of use	Pragmatic	Cognitive	Interview	Interview	[27]
Terminology	Terminology	Pragmatic	Process	QUIS (Questionnaire for User Interface Satisfaction) (0–9)	Questionnaire	[27]
Undefined	Immersion	Hedonic	Cognitive	Self-generated questionnaire	Questionnaire	[26]
	Understanding of the task and simplicity to manipulate	Hedonic	Process	Self-generated questionnaire	Questionnaire	[25]
	Graphic quality	Pragmatic	Process	Self-generated questionnaire	Questionnaire	[26]
	Motion sickness	Pragmatic	Physiological	Self-generated questionnaire	Questionnaire	[26]
	Learnability	Pragmatic	Cognitive	Self-generated questionnaire	Questionnaire	[25]
	Confidence	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[44]
	Enjoyment	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[44]
	Control of virtual objects	Hedonic	Process	Subjective judge (1–5 point scale)	Questionnaire	[44]
	Realism	Hedonic	Process	Subjective judge (1–5 point scale)	Questionnaire	[44]
Usability	Ease of movement	Pragmatic	Physiological	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Readability of the text	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Ability to control the machine	Pragmatic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Instructional understanding	Pragmatic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]

Table A2. *Cont.*

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
	Realism	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Ease of use	Pragmatic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Positive and negative comments	Hedonic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[36]
	Likeability	Hedonic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Complexity	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Ease of use	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Need of support	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Integration	Hedonic	Process	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Inconsistency	Pragmatic	Process	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Learnability	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Cumbersomeness	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Confidence	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Previous knowledge.	Pragmatic	Cognitive	SUS Scale (1–5)	Questionnaire	[27,34,35,46]
	Fixation	Pragmatic	Physiological	Eye tracking data analysis	Eye tracking data analysis	[26]
	Saccade	Pragmatic	Physiological	ND	Eye tracking data analysis	[26]
Usability	Visibility	Pragmatic	Physiological	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Similitude	Hedonic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	User control	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Consistency and standards	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Error prevention	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Preference	Pragmatic	Cognitive	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Flexibility	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Aesthetic properties	Hedonic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	User help	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Documentation	Pragmatic	Process	Self-generated survey/questionnaire (Likert scale)	Questionnaire	[26]
	Task execution time	Pragmatic	Process	Time	Manually-stopwatch	[26]

Table A2. *Cont.*

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Usability	Task adaptation time (time needed to adapt to the task before executing it)	Pragmatic	Process	Time	Manually-stopwatch	[41]
User Experience	Attractiveness	Hedonic	Cognitive	UEQ (User Experience Questionnaire Short Version (−3,3)	Questionnaire	[26]
	Efficiency	Pragmatic	Process	UEQ (User Experience Questionnaire Short Version (−3,3)	Questionnaire	[26]
	Comprehensibility	Pragmatic	Cognitive	UEQ (User Experience Questionnaire Short Version (−3,3)	Questionnaire	[26]
	Reliability	Pragmatic	Process	UEQ (User Experience Questionnaire Short Version (−3,3)	Questionnaire	[26]
	Stimulation	Pragmatic	Cognitive	UEQ (User Experience Questionnaire Short Version (−3,3)	Questionnaire	[26]
	Novelty	Hedonic	Cognitive	UEQ (User Experience Questionnaire Short Version (−3,3)	Questionnaire	[24]
	Amount of responsibility	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[28]
	Physical demand	Pragmatic	Physiological	Subjective judge (1–5 point scale)	Questionnaire	[28]
	Mental stress	Hedonic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[28]
	Attention required	Pragmatic	Cognitive	Subjective judge (1–5 point scale)	Questionnaire	[28]
	Interruptions or spare time	Pragmatic	Process	Subjective judge (1–5 point scale)	Questionnaire	[28]
User Participation	Brain effective connectivity	Pragmatic	Physiological	Rate of perceived exertion	fNIRS (Near-infrared spectroscopy) + polar watch	[32]
User Preference	Participant's overall preference	Hedonic	Cognitive	Subjective ranking of elements	Questionnaire	[46]
Workload	Fixation	Pragmatic	Physiological	ND	Eyetracking data analysis	[31]
	Saccade	Pragmatic	Physiological	ND	Eyetracking data analysis	[31]
	Mental demand	Pragmatic	Cognitive	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]
	Physical demand	Pragmatic	Physiological	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]
	Temporal demand	Pragmatic	Process	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]
	Performance	Pragmatic	Process	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]
	Effort	Pragmatic	Physiological	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]

Table A2. *Cont.*

Factor	Metrics	Hedonic vs. Pragmatic	Categories	Unit of Measurement	Technique/Data Collection Method	Paper
Workload	Frustration	Hedonic	Cognitive	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]
	Total workload	Pragmatic	Process	NASA-TLX Scale (1–10)	Questionnaire	[21,22,26,28, 31,37]
	Pupillometry	Pragmatic	Physiological	Pupil size variation/time	Eye tracking data analysis	[22]
	Perception of the visual feedback	Hedonic	Cognitive	Witmer and Singer presence questionnaire adaptation scale (1–10)	Questionnaire	[26]
	Perception of the auditory feedback	Hedonic	Cognitive	Witmer and Singer presence questionnaire adaptation scale (1–10)	Questionnaire	[26]
	Perception of the haptic feedback	Hedonic	Cognitive	Witmer and Singer presence questionnaire adaptation scale (1–10)	Questionnaire	[26]

References

1. Mourtzis, D.; Angelopoulos, J.; Panopoulos, N. A Literature Review of the Challenges and Opportunities of the Transition from Industry 4.0 to Society 5.0. *Energies* **2022**, *15*, 6276. [CrossRef]
2. Schroeter, R. Inception of Perception-Augmented Reality in Virtual Reality: Prototyping Human–Machine Interfaces for Automated Driving. In *User Experience Design in the Era of Automated Driving*; Springer International Publishing: Cham, Switzerland, 2022; Volume 980. [CrossRef]
3. Ivanov, D. The Industry 5.0 framework: Viability-based integration of the resilience, sustainability, and human-centricity perspectives. *Int. J. Prod. Res.* **2022**, *61*, 1683–1695. [CrossRef]
4. Lou, S.; Hu, Z.; Zhang, Y.; Feng, Y.; Zhou, M.C.; Lv, C. Human-Cyber-Physical System for Industry 5.0: A Review From a Human-Centric Perspective. *IEEE Trans. Autom. Sci. Eng.*. [CrossRef]
5. Grabowska, S.; Saniuk, S.; Gajdzik, B. Industry 5.0: Improving humanization and sustainability of Industry 4.0. *Scientometrics* **2022**, *127*, 3117–3144. [CrossRef] [PubMed]
6. Wang, J.; Lindeman, R. Coordinated 3D Interaction in Tablet- and HMD-BASED Hybrid Virtual Environments. In Proceedings of the SUI 2014-Proceedings of the 2nd ACM Symposium on Spatial User Interaction, New York, NY, USA, 4–5 October 2014; pp. 70–79. [CrossRef]
7. Forlizzi, J.; Battarbee, K. Understanding Experience in Interactive Systems. In Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods and Techniques, Cambridge, MA, USA, 1–4 August 2004; pp. 261–268. [CrossRef]
8. Hassenzahl, M.; Tractinsky, N. User experience-A research agenda. *Behav. Inf. Technol.* **2006**, *25*, 91–97. [CrossRef]
9. Kim, Y.M.; Rhiu, I.; Yun, M.H. A Systematic Review of a Virtual Reality System from the Perspective of User Experience. *Int. J. Hum. Comput. Interact.* **2019**, *36*, 893–910. [CrossRef]
10. Boletsis, C. The new era of virtual reality locomotion: A systematic literature review of techniques and a proposed typology. *Multimodal Technol. Interact.* **2017**, *1*, 24. [CrossRef]
11. Skarbez, R.; Smith, M.; Whitton, M.C. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Front. Virtual Real.* **2021**, *2*, 647997. [CrossRef]
12. Daling, L.M.; Schlittmeier, S.J. Effects of Augmented Reality-, Virtual Reality-, and Mixed Reality–Based Training on Objective Performance Measures and Subjective Evaluations in Manual Assembly Tasks: A Scoping Review. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2022**, *66*, 589–626. [CrossRef] [PubMed]
13. Stanney, K.M.; Mourant, R.R.; Kennedy, R.S. Human Factors Issues in Virtual Environments: A Review of the Literature. *Presence* **1998**, *7*, 327–351. [CrossRef]
14. Kaplan, A.D.; Cruit, J.; Endsley, M.; Beers, S.M.; Sawyer, B.D.; Hancock, P.A. The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis. *Hum. Factors* **2020**, *63*, 706–726. [CrossRef]

15. Kitchenham, B. Guidelines for Performing Systematic Literature Reviews in Software Engineering. 2007. Available online: https://www.researchgate.net/publication/302924724 (accessed on 5 January 2023).

16. Carrera-Rivera, A.; Ochoa, W.; Larrinaga, F.; Lasa, G. How-to conduct a systematic literature review: A quick guide for computer science research. *Comput. Ind.* **2022**, *142*, 101895. [CrossRef] [PubMed]

17. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, 71. [CrossRef] [PubMed]

18. Parsif.al. Available online: https://parsif.al/ (accessed on 25 September 2023).

19. Peruzzini, M.; Grandi, F.; Cavallaro, S.; Pellicciari, M. Using virtual manufacturing to design human-centric factories: An industrial case. *Int. J. Adv. Manuf. Technol.* **2020**, *115*, 873–887. [CrossRef]

20. Peruzzini, M.; Pellicciari, M.; Grandi, F.; Andrisano, A.O. Una configuración de realidad virtual multimodal para el diseño centrado en el ser humano de estaciones de trabajo industriales. *Dyna* **2019**, *94*, 182–188. [CrossRef]

21. Ahmed, S.; Demirel, H.O. A Framework to Assess Human Performance in Normal and Emergency Situations. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng.* **2020**, *6*, 011009. [CrossRef]

22. Nenna, F.; Orso, V.; Zanardi, D.; Gamberini, L. The virtualization of human–robot interactions: A user-centric workload assessment. *Virtual Real.* **2022**, *27*, 553–571. [CrossRef]

23. Ostrander, J.K.; Tucker, C.S.; Simpson, T.W.; Meisel, N.A. Evaluating the use of virtual reality to teach introductory concepts of additive manufacturing. *J. Mech. Des. Trans. ASME* **2019**, *142*, 051702. [CrossRef]

24. Pletz, C.; Zinn, B. Evaluation of an immersive virtual learning environment for operator training in mechanical and plant engineering using video analysis. *Br. J. Educ. Technol.* **2020**, *51*, 2159–2179. [CrossRef]

25. Hernández-Chávez, M.; Cortés-Caballero, J.M.; Pérez-Martínez, Á.A.; Hernández-Quintanar, L.F.; Roa-Tort, K.; Rivera-Fernández, J.D.; Fabila-Bustos, D.A. Development of virtual reality automotive lab for training in engineering students. *Sustainability* **2021**, *13*, 9776. [CrossRef]

26. Morosi, F.; Caruso, G. Configuring a VR simulator for the evaluation of advanced human–machine interfaces for hydraulic excavators. *Virtual Real.* **2021**, *26*, 801–816. [CrossRef]

27. Barot, C.; Lourdeaux, D.; Burkhardt, J.-M.; Amokrane, K.; Lenne, D. V3S: A Virtual Environment for Risk-Management Training Based on Human-Activity Models. *Presence Teleoperators Virtual Environ.* **2013**, *22*, 1–19. [CrossRef]

28. Khamaisi, R.K.; Brunzini, A.; Grandi, F.; Peruzzini, M.; Pellicciari, M. UX assessment strategy to identify potential stressful conditions for workers. *Robot. Comput. Integr. Manuf.* **2022**, *78*, 102403. [CrossRef]

29. Hoesterey, S.; Onnasch, L. A New Experimental Paradigm to Manipulate Risk in Human-Automation Research. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **2022**, *66*, 1170–1185. [CrossRef] [PubMed]

30. Kuts, V.; Marvel, J.A.; Aksu, M.; Pizzagalli, S.L.; Sarkans, M.; Bondarenko, Y.; Otto, T. Digital Twin as Industrial Robots Manipulation Validation Tool. *Robotics* **2022**, *11*, 113. [CrossRef]

31. Das, S.; Maiti, J.; Krishna, O. Assessing mental workload in virtual reality based EOT crane operations: A multi-measure approach. *Int. J. Ind. Ergon.* **2020**, *80*, 103017. [CrossRef]

32. Bu, L.; Chen, C.H.; Ng, K.K.H.; Zheng, P.; Dong, G.; Liu, H. A user-centric design approach for smart product-service systems using virtual reality: A case study. *J. Clean. Prod.* **2020**, *280*, 124413. [CrossRef]

33. Ricci, G.; De Crescenzio, F.; Santhosh, S.; Magosso, E.; Ursino, M. Relationship between electroencephalographic data and comfort perception captured in a Virtual Reality design environment of an aircraft cabin. *Sci. Rep.* **2022**, *12*, 10938. [CrossRef] [PubMed]

34. Doolani, S.; Owens, L.; Wessels, C.; Makedon, F. Vis: An immersive virtual storytelling system for vocational training. *Appl. Sci.* **2020**, *10*, 8143. [CrossRef]

35. Bernal, I.F.M.; Lozano-Ramírez, N.E.; Cortés, J.M.P.; Valdivia, S.; Muñoz, R.; Aragón, J.; García, R.; Hernández, G. An Immersive Virtual Reality Training Game for Power Substations Evaluated in Terms of Usability and Engagement. *Appl. Sci.* **2022**, *12*, 711. [CrossRef]

36. Rogers, C.B.; El-Mounayri, H.; Wasfy, T.; Satterwhite, J. Assessment of STEM e-Learning in an immersive virtual reality environment. *Comput. Educ. J.* **2017**, *8*, 1–12. [CrossRef]

37. Bernard, F.; Zare, M.; Sagot, J.C.; Paquin, R. Using Digital and Physical Simulation to Focus on Human Factors and Ergonomics in Aviation Maintainability. *Hum. Factors* **2019**, *62*, 37–54. [CrossRef] [PubMed]

38. Hart, S.; Stavenland, L. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Adv. Psychol.* **1988**, *52*, 139–183.

39. Chin, J.; Diehl, V.; Norman, K.L. Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. 1988. Available online: https://www.researchgate.net/publication/248594191 (accessed on 12 March 2023).

40. Schrepp, M. User Experience Questionnaire Handbook. 2023. Available online: www.ueq-online.org (accessed on 20 June 2023).

41. Torres, F.; Tovar, L.A.N.; del Rio, M.S. A learning evaluation for an immersive virtual laboratory for technical training applied into a welding workshop. *Eurasia J. Math. Sci. Technol. Educ.* **2017**, *13*, 521–532. [CrossRef]

42. Hassenzahl, M. The Thing and I: Understanding the Relationship Between User and Product. In *Funology: From Usability to Enjoyment*; Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003; Volume 2, pp. 31–42. [CrossRef]

43. Hassenzahl, M. The effect of perceived hedonic quality on product appealingness. *Int. J. Hum. Comput Interact.* **2001**, *13*, 481–499. [CrossRef]

44. Ho, N.; Wong, P.M.; Chua, M.; Chui, C.K. Virtual reality training for assembly of hybrid medical devices. *Multimed. Tools Appl.* **2018**, *77*, 30651–30682. [CrossRef]

45. Havard, V.; Jeanne, B.; Lacomblez, M.; Baudry, D. Digital twin and virtual reality: A co-simulation environment for design and assessment of industrial workstations. *Prod. Manuf. Res.* **2019**, *7*, 472–489. [CrossRef]

46. Zhang, L.; Liu, Y.; Bai, H.; Zou, Q.; Chang, Z.; He, W.; Wang, S.; Billinghurst, M. Robot-enabled tangible virtual assembly with coordinated midair object placement. *Robot. Comput. Manuf.* **2022**, *79*, 102434. [CrossRef]

47. Holmqvist, K.; Dewhurst, R.; van De Weijer, J. *Eye Tracking: A Comprehensive Guide To Methods And Measures*; OUP Oxford: Oxford, UK, 2011. Available online: https://www.researchgate.net/publication/254913339 (accessed on 10 May 2023).

48. Mosadeghi, S.; Reid, M.W.; Martinez, B.; Rosen, B.T.; Spiegel, B.M.R. Feasibility of an immersive virtual reality intervention for hospitalized patients: An observational cohort study. *JMIR Ment. Health* **2016**, *3*, e28. [CrossRef] [PubMed]

49. Young, C.B.; Mormino, E.C.; Poston, K.L.; Johnson, K.A.; Rentz, D.M.; Sperling, R.A.; Papp, K.V. Computerized cognitive practice effects in relation to amyloid and tau in preclinical Alzheimer's disease: Results from a multi-site cohort. *Alzheimer's Dement. Diagn. Assess. Dis. Monit.* **2023**, *15*, e12414. [CrossRef]

50. Sandelowski, M. Focus on research methods: Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Res. Nurs. Health* **2000**, *23*, 246–255. [CrossRef]

51. Apraiz, A.; Lasa, G.; Montagna, F.; Blandino, G.; Triviño-Tonato, E.; Dacal-Nieto, A. An Experimental Protocol for Human Stress Investigation in Manufacturing Contexts: Its Application in the NO-STRESS Project. *Systems* **2023**, *11*, 9. [CrossRef]

52. Paradis, E.; Sutkin, G. Beyond a good story: From Hawthorne Effect to reactivity in health professions education research. *Med. Educ.* **2017**, *51*, 31–39. [CrossRef]

53. Brown, P.; Spronck, P.; Powell, W. The simulator sickness questionnaire, and the erroneous zero baseline assumption. *Front. Virtual Real.* **2022**, *3*, 945800. [CrossRef]

54. Bouchard, S.; Berthiaume, M.; Robillard, G.; Forget, H.; Daudelin-Peltier, C.; Renaud, P.; Blais, C.; Fiset, D. Arguing in Favor of Revising the Simulator Sickness Questionnaire Factor Structure When Assessing Side Effects Induced by Immersions in Virtual Reality. *Front. Psychiatry* **2021**, *12*, 739742. [CrossRef]

55. Kim, H.K.; Park, J.; Choi, Y.; Choe, M. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Appl. Ergon.* **2018**, *69*, 66–73. [CrossRef] [PubMed]

56. Harris, D.; Wilson, M.; Vine, S. Development and validation of a simulation workload measure: The simulation task load index (SIM-TLX). *Virtual Real.* **2019**, *24*, 557–566. [CrossRef]

57. Dado, M.; Kotek, L.; Hnilica, R.; Tůma, Z. The Application of Virtual Reality for Hazard Identification Training in the Context of Ma-chinery Safety: A Preliminary Study. *Manuf. Technol.* **2018**, *18*, 732–736. [CrossRef]

58. Apple Inc. *Apple Vision Pro [Apparatus and Software]*; Apple Inc.: Cupertino, CA, USA, 2024.

59. Goodhue, D.L.; Thompson, R.L. Task-Technology Fit and Individual Performance. *MIS Q.* **1995**, *19*, 213–236. [CrossRef]

60. CRediT. Available online: https://credit.niso.org/ (accessed on 19 May 2024).

61. HTC Corporation. *HTC Vive [Apparatus and Software]*; HTC Corporation: Taoyuan, China, 2018.

62. HTC Corporation. *HTC Vive Pro [Apparatus and Software]*; HTC Corporation: Taoyuan, China, 2020.

63. HTC Corporation. *HTC Vive Pro Eye [Apparatus and Software]*; HTC Corporation: Taoyuan, China, 2019.

64. Facebook Technologies. *Oculus Quest 2 [Apparatus and Software]*; Facebook Technologies: Menlo Park, CA, USA, 2020.

65. Facebook Technologies. *Oculus Rift [Apparatus and Software]*; Facebook Technologies: Menlo Park, CA, USA, 2016.