



Evaluating the effectiveness of virtual reality for safety-relevant training: a systematic review

Hans Stefan¹ · Michael Mortimer¹ · Ben Horan¹

Received: 31 January 2023 / Accepted: 21 July 2023 / Published online: 19 August 2023
© The Author(s) 2023

Abstract

The commercial release of affordable, low-cost, and consumer-ready virtual reality (VR) devices has increased the accessibility for researchers to investigate the benefits of VR technology including those aimed at education and training. VR technology provides several opportunities that may provide benefits over traditional training methods, this is particularly relevant for safety training due to its ability to safely simulate dangerous scenarios that would otherwise be difficult to access. When implementing a new technology, it is important to evaluate and validate its effectiveness. This paper presents a systematic review of VR safety-relevant training studies that perform an evaluation of their effectiveness. This comprehensive review includes 136 studies published between 2016 and August 2021. Results presented in this paper include application domains, study objectives, study designs, and evaluation measures. Results show that the majority of studies were applicable to health services with the majority focusing on effectiveness evaluation using true- or quasi-experimental design. This study then categorizes each reported evaluation measure into one of the four levels in Kirkpatrick's model for training evaluation, results showed that the majority of studies evaluated learning (72.06%) and reaction (66.18%) levels with very few studies evaluating behavior and results levels. This study concludes by providing insights and recommendations to help future researchers make informed decisions when designing an effectiveness evaluation study for VR safety-relevant training applications.

Keywords Virtual reality · Evaluation methodologies · Safety training · Training evaluation · Training delivery method · Training effectiveness

1 Introduction

Safety training is widely recognized as an important means to reduce injuries, illnesses, and fatalities. Although safety training is considered to be beneficial, the degree of its effectiveness varies (Ricci et al. 2016; Robson et al. 2012). One of the main reasons for such variation is the different delivery methods used by training organizations (Burke et al. 2006). Although classroom training is the most commonly used method, it has been proven to be ineffective in achieving several desired outcomes such as safety knowledge

acquisition and improvements in attitudes, beliefs, behavior, and health (Ricci et al. 2016). Traditional delivery methods have been identified to possess several limitations, such as limited levels of engagement (Burke et al. 2006), difficulty in transferring the training to the real world (Gao et al. 2019), time inflexibility, and training inconsistencies due to instructor dependency. Besides safety training, these limitations can also be commonly generalized to other types of skills training that have indirect safety outcomes such as surgical, equipment assembly, operations, or maintenance training (i.e., safety-relevant training). With the majority of workplace accidents being considered preventable, continual improvements in safety and safety-relevant training are required to improve safety outcomes and reduce the number of workplace accidents. Virtual reality (VR) technology presents an important opportunity to improve the effectiveness of safety and safety-relevant training due to its increased level of presence, ability to fail safely, and capability in presenting several scenarios that are difficult to replicate in the real world due to financial constraints or safety concerns.

✉ Hans Stefan
h.stefan@deakin.edu.au
Michael Mortimer
m.mortimer@deakin.edu.au
Ben Horan
ben.horan@deakin.edu.au

¹ Deakin University, 75 Pigdons Rd, Waurn Ponds, VIC 3216, Australia

Advancements in VR technology and its availability, along with the realization of the opportunities that VR technology presents are among several factors that have contributed to the rapid increase in research related to the use of VR technology for safety-relevant training. Despite this increase, research focusing on how effectiveness evaluations for VR training are conducted remains limited. When evaluations are performed, the evidence supporting the effectiveness is either limited (Gao et al. 2019; Narciso et al. 2021) or inferior in quality, such as limited sample size or questionable research design (Jensen and Konradsen 2018; Renganayagalu et al. 2021; Tichon and Burgess-Limerick 2011). Although the consensus of the prospect of VR for training is positive, any perceived benefits need to be validated. Renganayagalu et al. (2021) presented an important finding that review studies investigating the effectiveness of VR for training are lacking. Before efforts are put toward evaluating the effectiveness of VR for training, it is important to first understand how these evaluations should be performed to ensure results are both valid and meaningful. This paper aims to understand how VR training evaluations are currently performed to provide insights into the design of VR training evaluations. These insights are expected to help guide future work in determining a more consistent and standardized approach for VR training evaluation. A more standardized approach is expected to lead to an increase in consistent and comparable research to better determine if VR is an effective tool for training and how it can be applied to safety-relevant training in general.

Although numerous opportunities for the use of VR technology in safety-relevant training have been identified, existing review studies targeted specific sectors, such as health services, construction, disaster preparedness, and mining. Other studies such as Renganayagalu et al. (2021) and Narciso et al. (2021) examined all professional training with the inclusion of safety-relevant training. In this case, safety was not discussed in isolation, making the value of VR for safety-relevant training difficult to determine. This paper aims to

extend existing research by reviewing evaluations used for safety-relevant training. This process is performed by first determining how VR safety-relevant training is evaluated and what methods are used (e.g., studies objectives, study designs, and application domains). This paper also analyses the evaluation measures used in VR safety-relevant training and categorizes those based on Kirkpatrick's four-level model to provide an overview of common measures used.

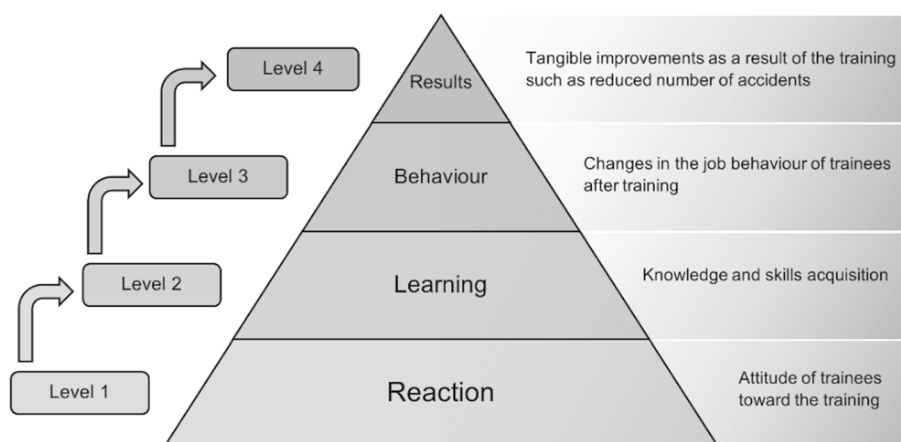
2 Background

This section provides a theoretical framework of training effectiveness evaluation based on Kirkpatrick's four-level model (Kirkpatrick 1976), followed by an overview of VR and safety-relevant training and related literature reviews.

2.1 Kirkpatrick's four-level model for training evaluation

The definition of effectiveness must be first described before performing evaluations. According to the Oxford Dictionary, effectiveness is “the degree to which something is successful in producing a desired result.” In the case of training effectiveness, the desired result differs depending on what is being trained. Kirkpatrick (1976) divided the process of evaluating the effectiveness of training into achievable measures in his four-level model (Fig. 1). In level 1, the measure of reaction gauges trainees' attitudes toward the training. In level 2, the measure of learning determines the level of knowledge and skills acquired. In level 3, the measure of behavior determines the level of change in trainees' related behavior after training. In Level 4, the measure of results relates to tangible improvements, such as a reduced number of safety incidents, increased revenue, or reduced operational costs. The Kirkpatrick model is hierarchical, indicating that the prior level is to be focused on and measured before progressing to a higher level of the model. This condition is

Fig. 1 Kirkpatrick's four-level model of training evaluation (Kirkpatrick 1976)



due to the concept that higher-level measurements are not expected to change if the lower levels of the model have not been satisfactory (Salas et al. 2012). For example, if trainees consider the training to be designed and delivered poorly, the training is unlikely to produce a meaningful increase in learning and unlikely to lead to behavioral changes or tangible results as a poor reaction tends to lead to trainee inattention and lack of engagement. According to Kirkpatrick and Kirkpatrick (2007, p. 123), although increases in the higher-level measurements are observed after training, evaluating levels sequentially is essential for building a compelling chain of evidence. When a chain of evidence is formed, the value of training becomes more meaningful (Kirkpatrick and Kirkpatrick 2007, p. 123). For example, an organization may implement new health and safety training with an observable reduction in accidents (level 4 of the Kirkpatrick model). This observable reduction in safety-related accidents can be due to a multitude of factors, with the introduction of new training being one possible factor. As the relationship between each factor and the observable result is indirect, it is difficult to ascertain which factors contributed to the reduction of accidents. Using Kirkpatrick's framework makes it easier to find these relationships between the training and a reduction in safety-related accidents by using the four levels of training evaluation to build a chain of evidence.

Several newer evaluation models have been proposed after the introduction of Kirkpatrick's model. Kraiger et al. (1993) suggested that learning outcomes are multidimensional, that is, learning should be evaluated by assessing the changes in trainees' cognitive, skill or affective capabilities. Kraiger et al. (1993) believed that Kirkpatrick's model is unclear in specifying these expected changes and their corresponding assessment technique. Kraiger et al. (1993) proposed a classification scheme to address this concern. Holton (1996) raised concerns with Kirkpatrick's model, particularly on the implied causality of the levels in the model and suggested that it is a taxonomy used beyond its scope rather than a model. Holton (1996) stressed the importance of intervening variables and identified three elements—ability, motivation, and environment—which can influence primary outcomes. Holton (1996) introduced three primary outcomes in his model: learning, individual performance, and organizational results. A model aimed at integrating multiple models together has also been established by Alvarez et al. (2004). This hybrid model combines four existing models proposed by Kirkpatrick (1976), Tannenbaum et al. (1993), Holton (1996), and Kraiger (2002). Alvarez et al. (2004) performed a review of empirical studies that analyzed the factors affecting the effectiveness of training and included the findings in their model. Although the hybrid model presented is more elaborate, Kirkpatrick's footprints are still evident considering that the other three included models are derived from Kirkpatrick's model.

Upon examining seven other training evaluation models (including Holton's (1996) model), Reio et al. (2017) found that most training evaluation models are variations of Kirkpatrick's four-level model. Despite its criticisms over the last five decades and the proposals of newer evaluation models, Kirkpatrick's model continues to be the most relevant and widely used model by different organizations as the basis for training evaluation (Bates 2004; Reio et al. 2017). The strength of Kirkpatrick's model comes from its simplicity and practicality (Reio et al. 2017), making it the most widely accepted and influential model (Phillips 2003). On the other hand, its criticism mainly stems from a lack of evidence in the sequential relationships of the levels (Alliger and Janak 1989). After careful consideration of the literature, Kirkpatrick's four-level model has been used in this review as the basis for categorizing evaluation measures of VR safety-relevant training. Kirkpatrick's levels were used solely for taxonomy, utilizing its applicability strength without focusing on the relationship between each level and their degree of importance, therefore, avoiding its major criticism.

2.2 VR and safety-relevant training

The term VR encompasses a wide range of technologies with one important characteristic, the ability to subject a user to an artificially generated environment. This range of technologies is often described by its level of immersion, which is the technical capabilities of a system to not only substitute real sensory information with computer-generated ones, but also to support natural actions for perceiving this information (Slater and Sanchez-Vives 2016; Slater 2009). For example, a highly immersive VR system such as the head-mounted display (HMD) allows its user to perceive visual and auditory (and sometimes tactile) information using natural actions such as reaching a hand to touch objects and turning the head to change viewpoint. In comparison, a low immersive VR system such as a desktop VR, despite also facilitating visual and auditory perceptions, requires unnatural actions such as using a computer mouse and keyboard to perceive the same sensory information. Another important concept in VR literature is presence, which is a user's subjective experience of being in the virtual environment (Schubert et al. 2001). Acknowledging that presence and immersion are sometimes used interchangeably or have different definitions (e.g., Witmer and Singer 1998), this paper follows the distinction between presence as a subjective experience and immersion as system characteristics as described previously.

The first wave of VR in the 1980s and 1990s created widespread public attention with many recognizing VR's potential to solve various real-world problems (Slater and Sanchez-Vives 2016). Researchers recognized the potential of VR in education and training with early applications

in aviation (Blake 1996; 1995; Page 2000), health (Satava 1995), and military sectors (National Research Council 1995; Hill et al. 2003). During this period, VR was still uncomfortable, experimental, and expensive, meaning that it was not viable for most users. As a result, attention from the public towards VR technology subsided, ending the first wave of VR. However, recent advancements in VR technology through the release of the Oculus Rift DK1 HMD in early 2013 marked what many refer to as the second wave of VR. This release was followed by the introduction of reliable, comfortable, and affordable commercial HMD devices, which helped bring comfortable and affordable VR technology to the masses. Besides the reduction in price, improvements can also be seen in the usability and ergonomics of VR hardware such as a reduction in weight and simpler setup, as well as in the quality of its output such as higher refresh rate, wider field of view, and introduction of haptic feedback.

This recent advancement in VR technology offers several opportunities that may help overcome current safety-relevant training limitations. For example, VR can simulate dangerous and difficult situations while still maintaining the safety of trainees (Czarnek et al. 2020; Freina and Ott 2015; Pedram et al. 2020); help motivate learners and keep them engaged with the learning activity (Freina and Ott 2015; Gao et al. 2019; Li et al. 2015; Sacks et al. 2013); and expose trainees to realistic simulated hazards which can improve the overall effectiveness of safety training (Burke et al. 2011). VR may also assist in transferring learning outcomes to the real world by recreating a learning context that is highly realistic (Freina and Ott 2015; Ganier et al. 2014; Rose et al. 2000). Another advantage of VR training is that it can be accessed on-demand, reducing potential scheduling issues and supporting just-in-time training. Finally, training can be delivered more consistently due to it being delivered by a system removing discrepancies that may exist between different trainers due to how they decide to deliver the content and other human factors.

2.3 Related literature reviews

Several related literature reviews have been identified that address the use of VR for learning and training. Abich et al. (2021) provided a domain-agnostic review to identify evidence of improvements in knowledge, skills, and abilities when using VR for training. The research included nine application domains in the review, these included aviation or aerospace, industry or manufacturing, military, first responder, general, medical, safety, education, and assembly. Radhakrishnan et al. (2021) performed a systematic review of immersive VR for skills training, which also included a wide range of industries. Suh and Prophet (2018) provided an overview of immersive technology research in various

areas, such as education, marketing, business, and health-care. Jensen and Konradsen (2018) investigated the use of VR for the education and training sector and narrowed the scope to only include HMDs. Apart from assessing the quality of the studies, Jensen and Konradsen (2018) emphasized how factors such as immersion, presence, physical discomfort, and attitude toward HMD technology influence learning outcomes (i.e., cognitive, affective, and psychomotor skills acquisition).

Other studies have focused on narrower scopes, these include education (Merchant et al. 2014; Radianti et al. 2020; Pellas et al. 2021) and training (Narciso et al. 2021; Renganayagalu et al. 2021). Education is the process of learning with the objective of acquiring knowledge, typically undertaken systematically in institutions such as schools and universities, whereas training refers to the process of learning with the objective of performing specific tasks or applying specific skills. Merchant et al. (2014) performed a meta-analysis examining the effectiveness of desktop-VR-based instructions with respect to learning outcomes for students in K-12 and higher education. Similarly, Pellas et al. (2021) also targeted K-12 and higher education although with a focus on immersive VR. The research by Radianti et al. (2020) complements Merchant et al. (2014) and Pellas et al. (2021)'s work by reviewing applications of immersive VR for higher education only. Narciso et al. (2021) synthesized the use of immersive VR for professional training focusing on the applied domain, hardware and stimuli used. The research provided a review of evaluation methods used by studies and the overall effectiveness of immersive VR for professional training. However, the results were only presented as complementary due to the fact that only 21 of the 66 studies included some form of evaluation. Renganayagalu et al. (2021) focused primarily on the effectiveness of VR HMDs for professional skill and safety training including only those studies that performed evaluations. Apart from discussing the effectiveness within different domains, the authors provided a summary of the types of skills trained, the training evaluation methods used, and the types of participants studied.

With regard to safety-relevant training, review studies typically focus on a specific field, making it difficult to identify benefits that may exist for all safety-relevant training. Figure 2 presents this context gap in systematic reviews with respect to safety-relevant VR training. As can be seen, the medical and surgical field has the highest number of reviews (12 review studies) as identified by a review conducted in 2020 (Abich et al. 2021). Laparoscopic surgery was the most popular application for VR surgical training (Gurusamy et al. 2008; Nagendran et al. 2013; Yiannakopoulou et al. 2015; Alaker et al. 2016). Other reviews focused on specific surgical procedures such as microsurgery (Erel et al. 2003), orthopedic surgery (Vaughan et al. 2016; Aïm et al. 2016), neurosurgery (Pelargos et al. 2017), and ear, nose, or throat

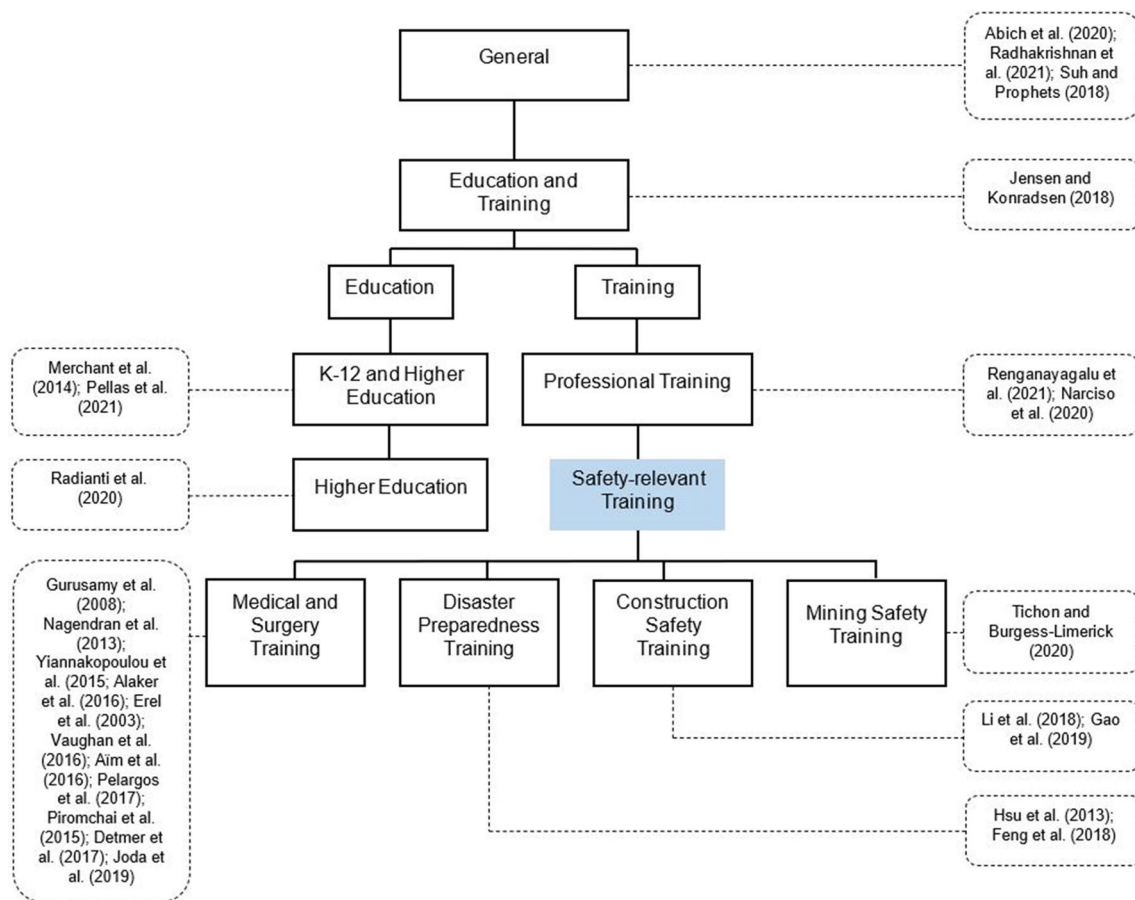


Fig. 2 Context gap

surgery (Piromchai et al. 2015). Other medical applications include renal interventions (Detmer et al. 2017) and dental medicine (Joda 2019). Besides the medical and surgical fields, other studies performed reviews for construction (Gao et al. 2019; Li et al. 2018), disaster preparedness (Hsu et al. 2013; Feng et al. 2018), and mining (Tichon and Burgess-Limerick 2011) sectors. Li et al. (2018) conducted a critical review of the applications of VR and augmented reality technologies in construction safety. Despite having a similar context, Gao et al. (2019) focused on training effectiveness rather than the application. The research investigated whether traditional training tools or computer-aided technologies are effective for acquiring knowledge, rectifying unsafe behavior, and reducing injuries. Feng et al. (2018) conducted a systematic review focusing on immersive VR serious games aimed at indoor evacuation processes. The research focused on the development and implementation criteria and generated a conceptual framework specific to evacuation. Hsu et al. (2013) reviewed several VR-based disaster preparedness and response training limited to the United States and presented the associated benefits and challenges of the implementation. Finally, Tichon and

Burgess-Limerick (2011) reviewed the use of VR for safety training in the mining sector. The research identified and briefly described a series of studies, identified trends and limitations within those studies, and drew conclusions and suggestions for future research directions.

3 Methods

The following sections describe the methods used for this systematic review, including the information sources, search strategy, selection process, and data collection process.

3.1 Information sources

The literature search process was performed during the month of August 2021 and covered four interdisciplinary databases (Scopus, Web of Science, EBSCOhost, and ACM digital library) and two discipline-specific databases (IEEE Xplore for computer science and engineering and PubMed for biomedicine and health).

3.2 Search strategy

The following search strategy was performed on the six databases using the search string below.

(“virtual reality” OR VR OR “virtual environment”) AND.

(evaluat* OR investigat* OR examin* OR assess* OR measur* OR compar*) AND.

(effect* OR impact OR outcome) AND.

(safe*) AND.

(train*) AND NOT.

(“machine learning” OR “deep learning”) AND NOT.

(“artificial intelligence” OR “neural network”) AND NOT.

(rehabilitation OR therapy).

This search string was used for initial screening from the title, abstract, and keywords field in Scopus (i.e., TITLE-ABS-KEY), topic field in Web of Science (i.e., TS), abstract field in EBSCOhost (i.e., AB), all field in PubMed, all meta-data field in IEEE Xplore, and abstract field in ACM digital library.

This paper focuses on studies that performed evaluations on VR training, therefore, six different synonyms of “evaluate” were used. An asterisk was used for each synonym to include different forms of the word such as “evaluate”, “evaluation”, and “evaluating”. Ensuring that these words were present helped limit the number of studies that focused on prototype development without performing any evaluation. Different words were also used to include the outcome of the evaluation such as “effectiveness”, “impact”, or “outcome”. The term “safe” and “train” were separated, and asterisks were used to include extended forms and combinations, including studies that did not explicitly phrase “safety training” together. Similar to Radianti et al. (2020), this review did not include the term “machine learning”, “deep learning”, “artificial intelligence”, or “neural network” to exclude studies that focused on artificial intelligence without the human learning context. This review also excluded the terms “rehabilitation” and “therapy” to remove studies that focused on personal health and fitness or focused solely on physical training with no learning.

3.3 Selection process

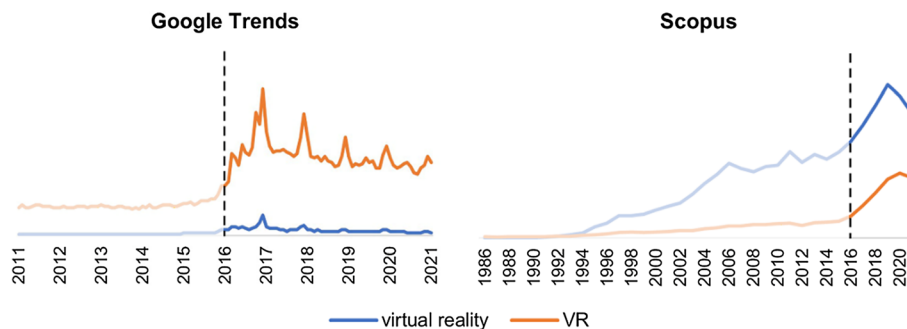
Studies identified throughout the database searching performed above were imported into a single EndNote Library. The authors then filtered out the studies prior to 2016 as it was considered to represent the start of a new wave of VR technology signified by the commercial release of VR HMDs (i.e., Oculus Rift in March 2016 and HTC VIVE in April 2016). This second wave of VR can be observed in an increased interest in VR technology from both public and academic domains. Figure 3 shows results from Google Trends and Scopus database that highlights this significant increase after 2016. The rationale for this research to focus on work conducted after 2016 is that significant advancements in VR technology at this time led to broader application and new research efforts. Although only papers published after 2016 were considered due to the release of significantly improved HMDs this work also includes desktop VR to ensure all levels of immersion are included for a comprehensive analysis.

After isolating the appropriate period, duplicate results were removed following the guidelines by Bramer et al. (2016). Primary sorting was then conducted by scanning the title and abstract of each study based on the following inclusion and exclusion criteria.

3.3.1 Inclusion criteria

Primary studies that evaluate the effectiveness of VR for safety or safety-relevant training. Safety training is defined as training with direct safety outcomes such as improvements in safety awareness, knowledge, or behavior and accident prevention. For example, hazard identification, electrical safety, and working at heights training are considered safety training. Safety-relevant training includes all types of skills training, where upon successful completion it will have a positive indirect effect towards the trainee’s own safety or safety of others. Examples of such training are surgical, equipment assembly, operations, or maintenance training.

Fig. 3 Google Trends searches worldwide and Scopus database number of documents by year for keywords “virtual reality” and “VR”



3.3.2 Exclusion criteria

- Nonempirical or secondary studies;
- Studies that are not published in English;
- Studies with unoriginal data;
- Studies that do not use or focus on VR technology;
- Studies that do not use VR for training;
- Studies with training unrelated to safety or focus on personal health (e.g., physical therapy, rehabilitation, or sport);
- Studies that do not evaluate the effectiveness of VR training;
- Studies or training intended for children.

This paper limits the scope of the review to studies intended for adults since adults learn differently compared to children (Knowles 1970) and therefore require separate analysis. The remaining articles that met the inclusion and exclusion criteria were then read in entirety to confirm eligibility using the same criteria described above. For articles that developed VR training without effectiveness evaluation but mentioned the intention in future research, efforts were made to identify these additional records. The selection and data collection processes were conducted primarily by the first author with a list of suggested exclusions presented to the second and third authors for articles that were ambiguous. The second and third authors then reviewed the suggested articles and either rejected or accepted the exclusion. Further ambiguity in selection and data collection was arbitrated by discussion. The first author read all selected studies at least twice, once for selection and then for coding. A flow diagram illustrating the process and the number of articles is presented in Fig. 4.

3.4 Data collection process

The following data was coded within a single spreadsheet after the study selection process was completed.

- Bibliographic information;
- Study purpose and training topic;
- Application domain;
- Studies design of the experiment;
- Whether the study focused on evaluation or development (study objective);
- Methods and measures used for evaluating the effectiveness.

3.4.1 Bibliographic information and study purpose

Bibliographic information such as the author(s), year published, and title, along with the study purpose and training topic were coded first as the process was simply a

duplication into the spreadsheet. The study purpose and training topic were obtained from explicit statements about both the aim and the training subject for each study. These were collected and used as references to the authors and were not used for any analysis.

3.4.2 Application domains

Application domain was coded primarily based on the list of industries and sectors described by the International Labor Organization (ILO; [n.d.a](#)) and the list of industries from WorkSafe Victoria (2022) with several additions from the authors. These additions were made to adjust the scope of the encompassing industry. For example, the term “maritime” or “offshore” safety was repeatedly used by studies and therefore was determined more suitable than “shipping; ports; fisheries” or “inland waterways” as described by the ILO. Another example is the military industry, which is specific enough to have a separate domain rather than to be included as part of “public service”. Other than these, there are other applications that were not described such as “general” and “space”.

3.4.3 Study designs

The study design was coded based on the methodological criteria for safety intervention evaluation research by Shannon et al. (1999). The study design was mainly categorized into true-experimental design, quasi-experimental design, and non-experimental design. A true-experimental design includes studies that have two or more independent groups where the participant allocation is randomized. When the allocation is not randomized, or the randomization is not mentioned, it is considered a quasi-experimental design. Non-experimental design includes studies with limited baseline comparison such as post-only one-group design, post-only non-equivalent control design, and pre-post one-group design (Shannon et al. 1999). True-experimental design is the ideal research design for maximizing internal validity. A quasi-experimental design is ideal where it may not be feasible to conduct a true experiment due to practical, ethical, political, or financial reasons (Shannon et al. 1999). Non-experimental design is considered to provide the least accurate results as it lacks a baseline measure used to compare intervention results, making the conclusions regarding its effectiveness less reliable (Shannon et al. 1999). As the name suggests, post-only one-group design involves collecting evaluation data after exposing all participants to VR training. Pre-post one-group design is similar to post-only one-group design with additional data collected before the training to measure the changes after training. Post-only non-equivalent control design is similar to a quasi-experimental between-groups design, without controlling the baseline of

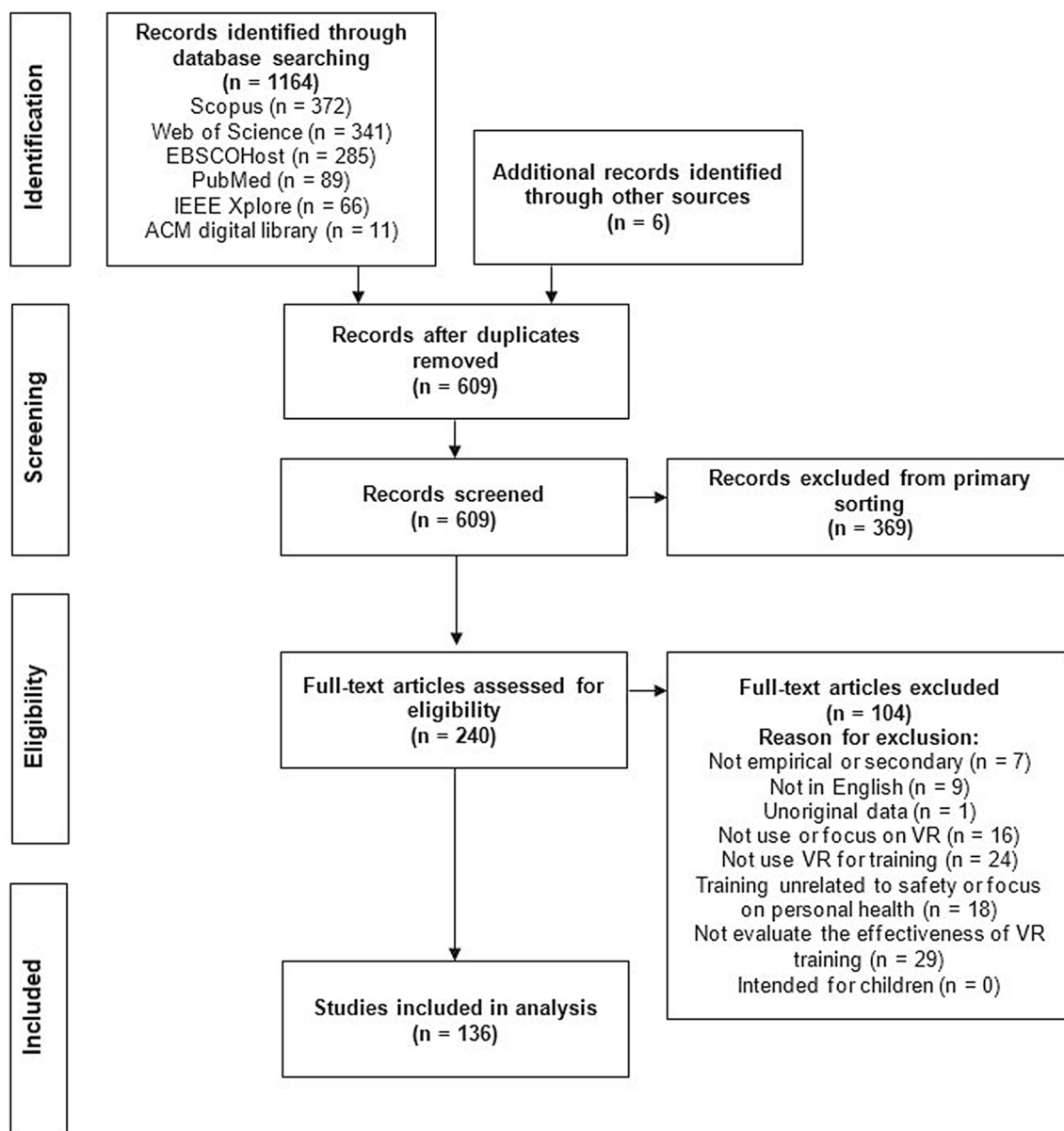


Fig. 4 PRISMA flow diagram of the review process

each participant that is normally captured using a pretest or a demographic survey. Besides this, for true-experimental and quasi-experimental design, each study is also categorized into between-groups, within-groups, or mixed design. Between-group design involves studies comparing dependent variables or measures between two independent groups. In the criteria explained by Shannon et al. (1999), within-group design, as part of the quasi-experimental design is limited to interrupted time-series design, which requires multiple data collection points both before and after participants undertake training. This review adopts a more general within-group design definition, which is studies that have all participants experiencing all conditions. These studies

include evaluation, where two (or more) training conditions (e.g., traditional and VR training) are experienced by all participants and data, is collected either after each exposure (i.e., learning curve) or once at the end. A mixed design incorporates studies that have a combination of between-groups and within-groups designs.

3.4.4 Study objectives

The study objective was coded based on two categories namely, evaluation and development. Evaluation refers to studies with the main objective of evaluating the effectiveness of VR training, whether it is an existing or a newly

developed prototype. A study was deemed to have evaluation as the study objective when it was stated as the aim of the study and where the primary focus of the results section was on presenting and discussing evaluation results. Development refers to studies where the main objective of the study was describing the software or hardware development of a VR training prototype with less emphasis on evaluation. This reduced emphasis was noted when the majority of the results were related to development with a separate and smaller section for evaluation of the training. The aim of defining studies using these categories is to provide insights into the research motivations and to distinguish the level of focus on improving training results. For example, a development study may focus on developing a working prototype to validate technological outcomes before improving training elements.

3.4.5 Evaluation measures

Reported measures for each study were recorded comprehensively based on the description given by the authors. The descriptions of measures for each of the studies were then grouped into categories of similar measures. The categories were then coded into one of the four levels of Kirkpatrick’s model (reaction, learning, behavior, and results). A measure is categorized as reaction if it is a subjective impression of the training obtained from either surveys, comments, discussions, or interviews. A measure is categorized as learning if an objective test for either knowledge or skills was conducted. A measure is categorized as behavior if

the behavior of participants was observed after the training during a period of work experience. A measure is categorized as results if it measures tangible improvements in the form of quantitative data after the training such as human resources data on reduction in the number of accidents. In the case of studies that reported more than one experiment in a single paper, data collected on each experiment was weighted appropriately based on the number of experiments. For example, papers reporting two experiments (Chen et al. 2020; Liang et al. 2019b; Clifford et al. 2019; Mirauda et al. 2020; Leder et al. 2019; Zhang et al. 2017) were weighted by multiplying 0.5 to the coded data for each of the experiments. The purpose of this weighting is to maintain an equal value for each paper.

4 Results and discussion

This section presents the results and discussion of the systematic review to provide an overview of how VR safety-relevant training is currently being evaluated for its effectiveness (Sects. 4.3) with an analysis of evaluation measures used, which are then categorized into the four levels of Kirkpatrick’s model (Sect. 4.4).

4.1 Application domain

Figure 5 presents an overview of the number of studies in different application domains for VR safety-relevant

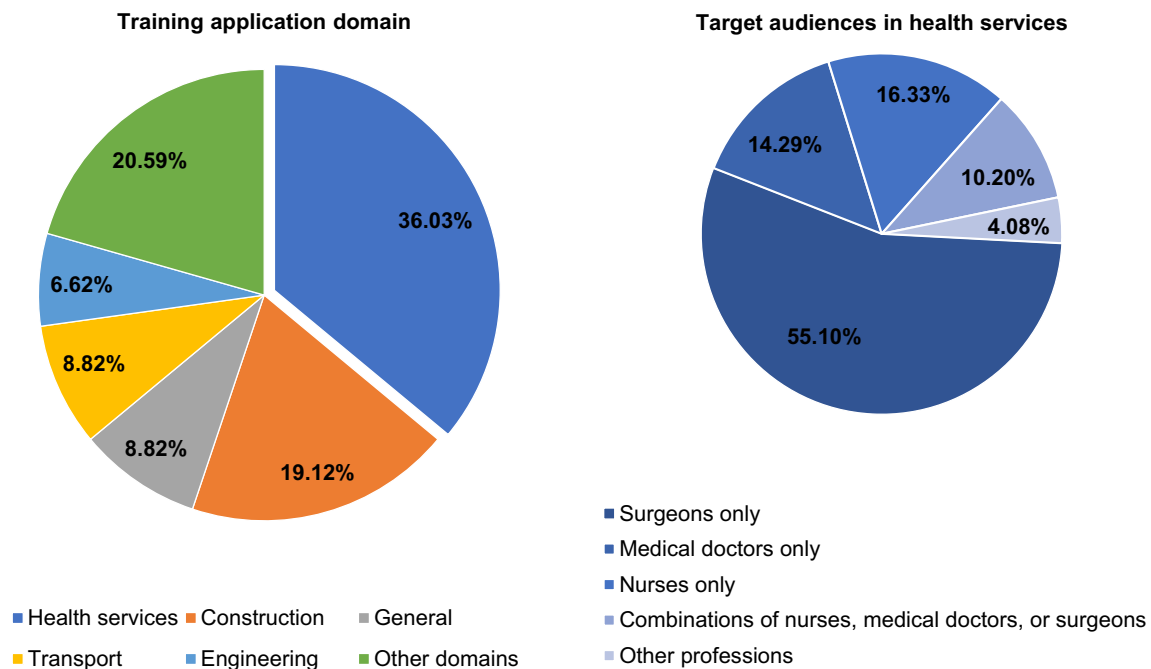


Fig. 5 Training application domain out of all 136 studies (left) and target audiences of VR training out of 49 studies in health services (right)

training. Health services were the most active domain, consisting of 36.03% of all studies reviewed, the breakdown of the target audiences within this domain is also presented in Fig. 5. Construction was the second most active domain with 19.12% of all studies. Studies without a specific application domain (e.g., disaster preparedness or slip, trip, and fall prevention training) were categorized as “general” and represent 8.82% of all studies, another 8.82% of all studies were conducted within the transportation domain, and 6.62% of all studies were related to engineering. Another 9 application domains represented the remaining 20.59% of all studies and were categorized under “other domains”, each of these domains comprises 5 studies or less. Application domains represented in the “other domains” category included utilities, manufacturing, maritime or offshore, education, automotive, emergency services, mining, military, and space.

4.1.1 Health services

Within health services, three key professions had a strong focus on VR training, this included surgeons, medical doctors, and nurses. More than half of studies within health services (55.10%) are aimed specifically toward surgeons, as can be observed on the right chart of Fig. 5. This is expected as surgery training is one of the earliest adopters of VR, and the abundance of commercially available VR simulators for surgery has resulted in a significant number of studies evaluating and validating the effectiveness. For example, dV-Trainer (Mimic Technologies Inc, Seattle, WA) has been frequently used and studied as a simulator for training robot-assisted surgery using the da Vinci Surgical System (Intuitive Surgical Inc, Sunnyvale, CA). Furthermore, since surgery is a highly dangerous procedure, competence must be confidently ensured. To do so, rigorous evaluations are conducted to validate the effectiveness of VR training to confirm its ability to accurately measure level of competency. Studies aimed at medical doctors (excluding surgeons) represented 14.29% of all studies in health services, this includes general practitioners and medical specialists, such as neonatologists (Xiao et al. 2020), anesthesiologists (Casso et al. 2019; Shewaga et al. 2018), and cardiologists (Jensen et al. 2016). Studies aimed at nurses were similarly represented, consisting of 16.33% of all health services studies. VR training aimed at nurses is relatively diverse and ranges from direct safety training (e.g., operating room fire, infection control, and home hazards training; (Polivka et al. 2019; Rossler et al. 2019; Yu et al. 2021)), to procedural safety-relevant training (e.g., catheter insertion, chemotherapy administration, childbirth, surgery assistance, and nasal sample collection training; (Butt et al. 2018; Cecil et al. 2021; Chan et al. 2021; Chang et al. 2019; Edwards et al. 2021)). Other professions in health services included

general healthcare staff (Rahouti et al. 2021) and laboratory technologists (Prendinger et al. 2016).

4.1.2 Construction

The construction domain also includes a substantial number of VR safety training studies with the majority of those studies focusing on general occupational health and safety and hazard identification and management training. In addition to a strong focus on typical hazards found on construction sites, a smaller number of studies focused on specific hazards and machinery, such as demolition robots (Adami et al. 2021), cranes (Dhalmahapatra et al. 2021; Song et al. 2021), precast/prestressed concrete (Joshi et al. 2021), and rollers (Vahdatikhaki et al. 2019). The focus on improving the safety of construction workers is expected to continue as fatality rates in construction have remained relatively consistent over the past 5 years with 3.2 and 3.1 deaths per 100,000 workers in Australia during 2015 and 2020 respectively (Safe Work Australia 2016a, 2021) and 10.1 and 10.2 in the USA during 2015 and 2020 respectively (Bureau of Labor Statistics 2016, 2021). Construction represented the third highest fatality rate in Australia in 2020 behind agriculture, fishing, and forestry; and transport, postal, and warehousing (Safe Work Australia 2021). The high number of studies that focus on evaluating the effectiveness of VR training applications in the construction domain suggests that improvement is a strong focus, with researchers attempting to use innovative technology to improve safe behavior to reduce accidents. However, unlike VR surgical simulators, VR safety-relevant training in the construction domain is largely one-off experimental prototypes, often lacking in follow-up improvements and evaluations or actual implementations. This is perhaps due to the difference in the level of incentive to implement VR training between the construction and health services domain such as surgery training. In construction, traditional training is relatively easy and inexpensive to deliver, with the training typically being conducted by outsourcing the training to a third party (i.e., training providers). There are often multiple training providers available in an area providing compliance training of similar quality with competitive fees. In comparison, the need for effective VR surgery simulators is greater due to the expensive and logistically demanding nature of traditional surgery training. Expert surgeons able to assist in training are less available, and traditional surgery training typically involves the elaborate use of cadavers or expensive realistic models. Although the implementation rate of VR safety-relevant training may be slower for the construction domain compared to surgery training, if the validity and effectiveness of VR safety-relevant training in construction is increasingly established, it is likely to increase the demand and development and led to commercially viable VR training

systems rather than prototype solutions. Recently, attempts have been made to commercialize VR training in the construction domain such as by Pixa VR (n.d.a) and Next World Enterprises (2022). Future research should focus on continually evaluating the effectiveness of VR training in the construction domain with a focus on implementing such training into organization training practices to progress from the current experimental prototypes to a more widely adopted and commercially viable training system.

4.1.3 General

The category “general” presented in Fig. 5 (left) represents VR training for which the application is broader than a single application domain. Examples included studies that focused on general public training on preparing for disasters such as fire (Benvegnù et al. 2021; Fu and Li 2020; Liang et al. 2020, 2019a; Lovreglio et al. 2021; Saghafian et al. 2020), earthquake (Li et al. 2017; Liang et al. 2018), and active shooters (Sharma et al. 2020); slip, trip, and fall prevention training (LoJacono et al. 2018; Weber, et al. 2020) and general indoor hazard awareness and response training (Cavalcanti et al. 2021).

4.1.4 Transport

Transport safety training studies focus on road (6 studies), civil aviation (5 studies), and railway safety (1 study). VR training for improving road safety is typically aimed at training regular drivers’ risk awareness and safe driving habits (Agrawal et al. 2017; Lang et al. 2018; Suto et al. 2020). Other road safety studies focus on efficient paths for performance driver training (Simpson and Rafferty 2020), interactions with automated cars (Sportillo et al. 2018), and bicycle safety training (Tsuboi et al. 2018). Four of the five studies that focus on civil aviation, trained passengers and flight crews on emergency-related procedures. The remaining study focuses on railway safety with respect to crane operations and signalling training for clearing railways after an accident (Xu et al. 2019b).

4.1.5 Engineering

The engineering domain includes studies that are applicable to all fields of engineering, such as job safety analysis training (Ta et al. 2019), workshop and laboratory safety training (Makransky et al. 2019; Mondragón-Bernal 2020; Sim et al. 2019), and discipline-specific training aimed at chemical (Chen et al. 2020; Colombo and Golzio 2016; Ouyang et al. 2018), mechanical (Keßler et al. 2020), and gas engineers (Asghar et al. 2019).

4.1.6 Other domains

A strong focus on VR training applications for safety-relevant training is prominent within safety-critical fields, such as health services, construction, transport, and engineering. Although not as strong, interest in VR training also exists in other safety-critical domains, such as utilities (Avveduto et al. 2017; García et al. 2016; Herrington and Tacy 2020; Kwegyir-Afful and Kantola 2021; Mirauda et al. 2020), manufacturing (Caporusso et al. 2019; Lacko 2020; Leder et al. 2019; Torres-Guerrero et al. 2019), maritime or offshore (Chae et al. 2021; Jung and Ahn 2018; Jung and Kim 2017; Smith et al. 2019; Smith and Veitch 2019), automotive (Borsci et al. 2016; Cooper et al. 2021; Sebastian et al. 2018), emergency services (Chen et al. 2021; Druzhinina et al. 2019; Prasolova-Førland et al. 2017), mining (Liang et al. 2019b; Zhang 2017), military (Clifford et al. 2019; Salcedo et al. 2016), and space (Liu et al. 2016). These findings are consistent with the need for safety-critical fields to always improve safety and prevent accidents with effective training considered an important factor in reducing safety incidents.

4.1.7 Agriculture, fishing, and forestry

It is interesting to note that no studies were found that have a focus on VR training for the agriculture, fishing, and forestry industry. This is unexpected considering this industry had the highest fatality rates for 2015 both in the USA (Bureau of Labor Statistics 2016) and Australia (Safe Work Australia 2016a) with 22.8 and 16.7 fatal injuries per 100,000 workers respectively. Although these fatality rates are decreasing, it remains the industry with the highest fatality rate in 2020 with a rate of 21.5 and 13.1 in the USA and Australia respectively (Bureau of Labor Statistics 2021; Safe Work Australia 2021). One possible factor for the lack of VR safety training in this domain could be the lack of safety training applied in general. The agriculture, fishing and forestry domain is unique in that a large proportion of the workforce are self-employed (Chapman and Husberg 2008; Safe Work Australia 2016b) without access to dedicated safety officers (McBain-Rigg et al. 2017) making it more difficult to ensure safety compliance when compared to other domains where the majority of the workforce is directly employed with an organization. The high fatality rate and the inability to reduce accidents using conventional methods demonstrate that the agriculture, fishing and forestry domain is still lagging behind other domains when it comes to the implementation of occupational health and safety. This lack of focus has likely resulted in a low motivation to continually improve and therefore there is a lack of desire to use innovative technologies such as VR. However, the shortcomings of conventional methods may provide an opportunity

for alternative methods such as VR training as a potential training solution. The majority of fatalities in the agriculture, fishing and forestry domain involved the (mis)use of vehicles such as tractors (Safe Work Australia 2016b). VR training may provide a safe alternative to train employees on how to safely operate such vehicles without exposing them to physical risk. This can be done through the use of vehicle simulation as has been done for other types of vehicles such as those solutions in the transportation domain. Furthermore, the recent availability of mobile and portable VR hardware may also provide a greater opportunity to deliver safety-related training in rural and remote areas where work in this domain is commonly performed. Future researchers are encouraged to develop and evaluate the effectiveness of VR training applicable to the agriculture, fishing, and forestry domain.

4.2 Study objectives

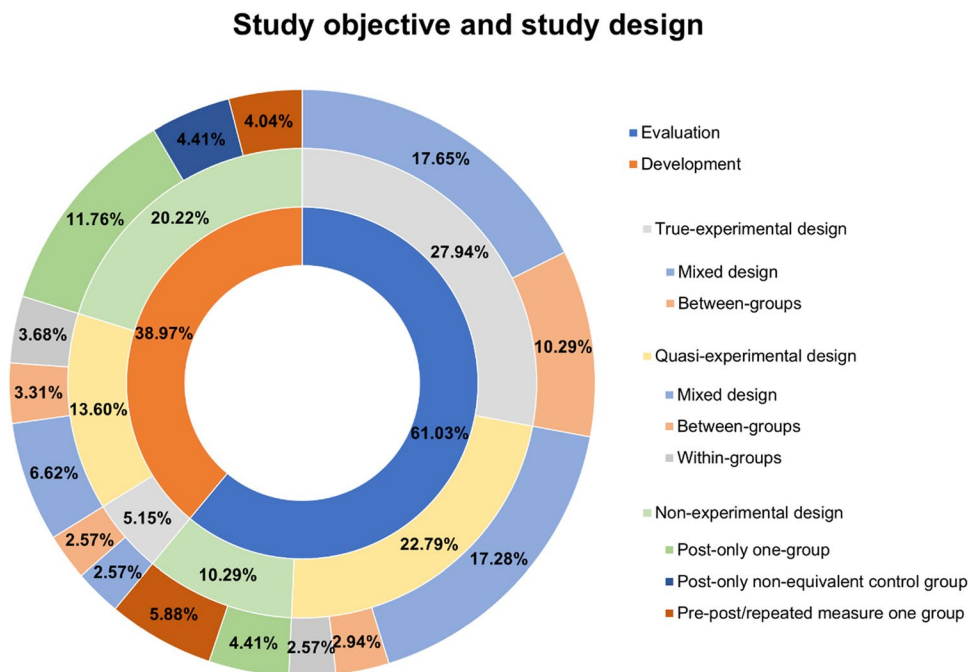
This review categorizes papers into two main study objectives, evaluation and development. As illustrated in Fig. 6, 61.03% of the studies aimed to evaluate the effectiveness of VR training. Effectiveness was typically evaluated directly by evaluating the effectiveness measures of the overall VR training either before and after training or compared against the control group (refer to study designs below). Besides this direct evaluation, other studies focused on specific factors expected to influence the effectiveness of training (i.e., indirect). Examples of factors that researchers highlighted, which indirectly influenced effectiveness (i.e., indirect evaluation), were the immersion of the input and output devices,

different learning strategies, anxiety levels (Kwon 2020), confidence levels, and prior experiences (Truong et al. 2021). Studies that investigated how the level of immersion can impact VR training effectiveness evaluated the use of different interfaces and display types, such as desktop or HMD (Buttussi and Chittaro 2018; Jung and Ahn 2018), virtually generated or 360° imagery (Moore et al. 2019), stereoscopic display (Tawadrous et al. 2017), and different spatial references (Fu and Li 2020); different controls (e.g., keyboard and mouse, controllers, hand tracking or no control (Burigat and Chittaro 2016)); and presence of additional multisensory stimuli, including haptic feedback (Cooper et al. 2021; Francone et al. 2019; Kim et al. 2020; Simpson and Rafferty 2020). The influences of learning strategies in VR training were evaluated in terms of utilizing gamification (Cavalcanti et al. 2021), presenting safety instructions with positive or negative consequences (Shi et al. 2019), lecture-based or mastery learning (Smith and Veitch 2019), and other strategies (Orland et al. 2020; Salcedo et al. 2016; Sebastian et al. 2018). The remaining 38.97% of studies focus on describing the development of VR training with supplementary effectiveness evaluation.

4.3 Study designs

As observed in Fig. 6, the majority of studies with evaluation as the main objective performed either a true experiment or a quasi-experiment representing 27.94% and 22.79% of all studies respectively and 83.12% of “evaluation” based studies used either true experimental or quasi-experimental design. In contrast, over half of the “development” studies

Fig. 6 Percentage study objective overall (inner circle) and percentage study designs by study objective (two outer circles)



used a non-experimental design with post-only one-group design being the most common (i.e., 11.76% of all studies). It is suggested that a non-experimental design is preferred for development studies due to its ability to quickly and easily identify quantifiable values of development, such as using simple usability or user experience questionnaires. The facts that the majority of studies focus on evaluation (i.e., 61.03%) and most of these studies utilize good study designs (i.e., 83.12%) are encouraging in terms of the state of VR for safety-relevant training. This is because researchers often focus on development studies when the technology is relatively new, and only when the technology is more well-understood do researchers attempt to evaluate how effective it is. Although development studies are necessary to determine and advance the capabilities of VR technology, studies aiming to understand how effective VR training is and how VR training could be implemented effectively may now benefit from a shift towards higher priority. This is likely due to the fact that VR technology has seen considerable technological advancements to the point that the technology has matured and can now provide highly functional, reliable, and usable training experiences. A shift in priority towards evaluation studies helps validate if and where VR training is effective for safety-relevant training in general and more importantly understand the role VR technology could play in safety-relevant training. As with the different training methods available (e.g., traditional classroom, computer-based/e-learning, and hands-on), understanding which types of safety-relevant training are best suited to take advantage of VR is worthwhile to investigate. This could help determine if current safety training could be improved through the use of this relatively new technology. Despite the importance of evaluation, development studies play a key role in advancing the current state of VR technologies. When development studies are undertaken, it is recommended that such studies consider validating their VR training by employing either a true- or quasi-experimental design as opposed to a simple post-only one-group design as is typically done. This review encourages future researchers, particularly those undertaking development studies, to consider improving their experimental design to increase the quality of their results. For example, instead of a post-only one-group or pre-post one-group design, researchers may benefit from multiple data collection points before and after training and move away from a non-experimental design to a quasi-experimental interrupted time series design. Another suggestion is for researchers to expose all participants to a control such as traditional training in addition to the VR training (i.e., within-groups design), and capture the outcome measure after each exposure, preferably allowing considerable interval between each exposure. Finally, a post-only non-equivalent control group may be enhanced by controlling a group baseline using a pretest.

With respect to true- and quasi-experimental design, a mixed design is the most common approach used by both “evaluation” and “development” studies, followed by between-groups and then within-groups design. Overall, mixed design covers 44.12% of all studies with 20.22% being a true experiment and 23.90% a quasi-experiment. Between-groups design represents 19.11% of all studies with 12.86% representing true experiment and 6.25% a quasi-experiment, whereas within-groups design only represents 6.25% of all studies. While there is nothing inherently good or bad about each of the design approaches, they do have their own advantages and disadvantages. Therefore, the correct use of these designs is best determined on a case-by-case basis, which is out of the scope of this review. This review instead presents the findings on study design as an overview of the current study design trend seen by research conducted in the use of VR for safety-relevant training.

When analyzing study design based on the application domain, VR training in health services represented a large portion of true- and quasi-experimental design with 13.97% and 12.50% respectively (see Fig. 7). This supports the previous rationale (Sect. 4.1.1) that existing commercially available VR simulators and the dangerous nature of this domain, particularly for surgeons, require rigorous testing with reliable experimental design. Other than health services, transportation is the only other application domain where non-experimental design is the least used approach. The majority of studies in construction, engineering, and other domains adopted either non- or quasi-experimental design with a smaller portion adopting true experiments. Less rigorous evaluation in “other domains” is reasonable considering the utilization of VR for training in these fields is still uncommon and largely exploratory. However, it was interesting to find that the construction domain, with the second largest application, still had a limited number of studies utilizing true-experimental design. Researchers conducting studies in the construction domain are encouraged to continue undertaking evaluative studies prioritizing a more comprehensive experimental design approach (i.e., true or quasi-experimental).

4.4 Evaluation measures categorized using Kirkpatrick’s four level model

This section presents an analysis of the evaluation measures used by reviewed studies which are categorized into the four levels in Kirkpatrick’s model as presented in Fig. 8. It is important to note that some of the reviewed studies evaluate more than one measure and therefore may be represented in multiple categories. Learning (level 2 in Kirkpatrick’s model) was the most commonly used evaluation measure with 98 (72.06%) studies. This was closely followed by studies that measured reaction (level 1 in Kirkpatrick’s model)

Fig. 7 Percentage study design based on application domain

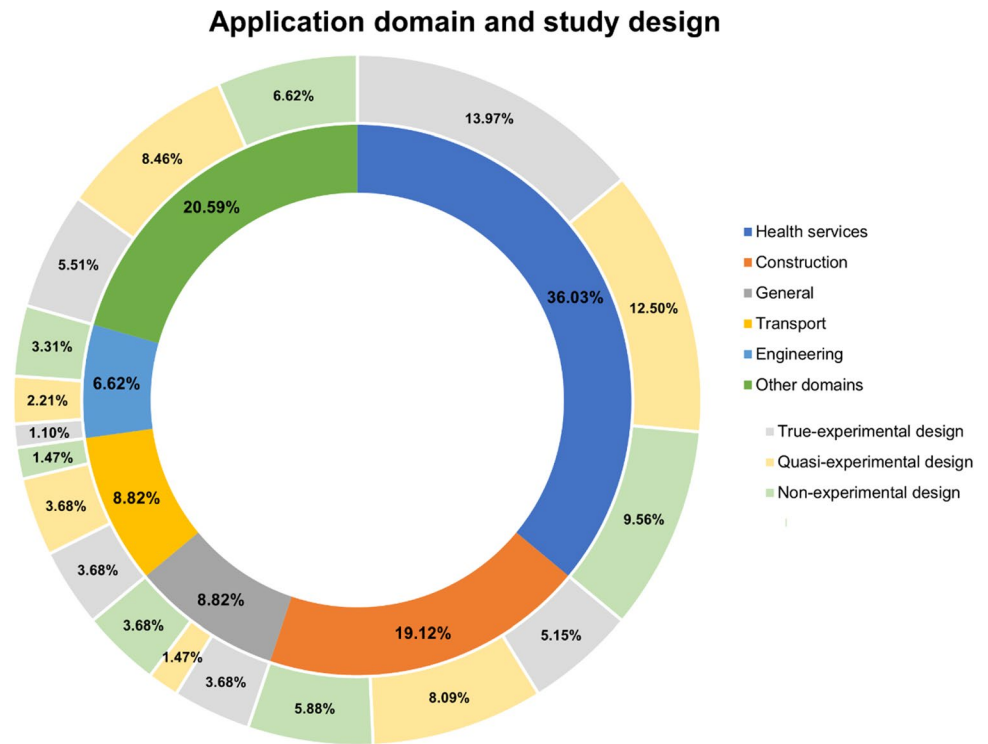
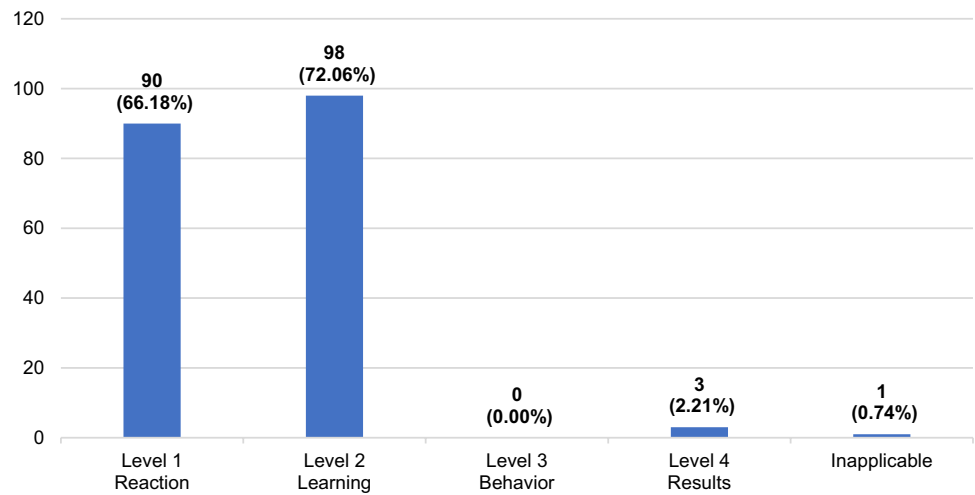


Fig. 8 Evaluation measures categorized into Kirkpatrick’s four level model



with 90 (66.18%) studies. A relatively low number of studies evaluated the effectiveness of VR training using behavior or result measures, which are beyond level 2 of Kirkpatrick’s model. In fact, no study reviewed evaluated behavior (level 3 of Kirkpatrick’s model), and only three studies evaluated results (level 4 of Kirkpatrick’s model). One study had measures that are inapplicable to any of the levels in this model (Torres-Guerrero et al. 2019). Torres-Guerrero et al. (2019) collected electroencephalogram (EEG) signals to measure the stress and concentration level when performing a welding task.

4.4.1 Level 1—reaction

Studies evaluating reaction typically used a post-survey in the form of a Likert scale or open-ended questionnaires to measure reaction, the surveys were either self-developed or based on existing work. Reaction was also measured using interviews, discussions, and observations. As presented in Table 1, reaction is categorized into 12 measures for the purpose of this review, namely affective reaction, realism, quality of instruction and feedback, motivation, usability, perceived learning effectiveness, subjective comparison, presence, engagement and interactiveness, intention to use,

Table 1 Categorized reaction measures and associated studies

Categorized measure	No. of studies (%)	Raw measures	Studies
Affective reaction	28 (20.59%)	General experience; Overall feeling; Positive response for use of VR system for training; Felt safe; Attitude towards training subject; Perceived enjoyment; Pleasantness; Attractiveness; Impressiveness; Desirability; Satisfaction; Emotional experience; Feeling towards the VR training	Ahn et al. (2020), Benda et al. (2020), Burigat and Chittaro (2016), Butt et al. (2018), Buttussi and Chittaro (2021), Cavalcanti et al. (2021), Chae et al. (2021), Chan et al. (2021), Chang et al. (2019), Druzhinina et al. (2019), Dzeng et al. (2016), Gupta and Varghese (2020), Chen et al. (2021), Herrington and Tacy (2020), Jung and Kim (2017), Makransky et al. (2019), Mondragón-Bernal (2020), Nykänen et al. (2020), Oh et al. (2020), Polivka et al. (2019), Chen et al. (2020), Rahouti et al. (2021), Xu et al. (2019a), Saghafian et al. (2020), Sim et al. (2019), Sportillo et al. (2018), Xu and Zheng (2021), Yu et al. (2021), Liang et al. (2019b)
Realism	21.5 (15.81%)	Realism; Representational fidelity; Image quality	Ahn et al. (2020), Bunes and Karlisen (2019), Casso et al. (2019), Chae et al. (2021), Di Vecce et al. (2021), Dorozhkin et al. (2017), Chen et al. (2021), Hoogenes et al. (2018), Kwegvir-Afful and Kantola (2021), Li et al. (2017), Mirauda et al. (2020), Prasolova-Førland et al. (2017), Rahouti et al. (2021), Saghafian et al. (2020), Sankaranarayanan et al. (2016), Schulz et al. (2019), Shewaga et al. (2019), Sportillo et al. (2018), Truong et al. (2021), Tsuboi et al. (2018), Xiao et al. (2020), Xu and Zheng (2021)
Quality of instruction or feedback	12 (8.82%)	Perception of content; Perception of guidance; Content design; Content ease to follow; Pertinence of the questions to the content taught; Difficulty of level change; Recommendations efficacy; Recommendations simplicity; Feedback quality; Instruction simplicity	Avveduto et al. (2017), Burigat and Chittaro (2016), Chae et al. (2021), Chan et al. (2021), Chittaro et al. (2018), Dzeng et al. (2016), Chen et al. (2021), Lovreglio et al. (2021), Mirauda et al. (2020), Chen et al. (2020), Rahouti et al. (2021), Tang et al. (2021), Xiao et al. (2020)
Motivation	8 (5.88%)	Motivation; Intrinsic motivation; Safety motivation	Chang et al. (2019), Dzeng et al. (2016), Joshi et al. (2021), Makransky et al. (2019), Nykänen et al. (2020), Pham et al. (2018), Rahouti et al. (2021), Tang et al. (2021)

Table 1 (continued)

Categorized measure	No. of studies (%)	Raw measures	Studies
Usability	48.5 (35.66%)	Usability; Perceived ease of use; Equipment understanding; Handling; System reactivity; Helpfulness of features; Quality and usefulness of force feedback; User-friendliness; Navigation comfort; Convenience; Accessibility; Motion consistency of tools; Acceptance of visualisation; Intuitiveness; Difficulty to understand; Difficulty to use; Additional difficulties; Confusion; Annoyance; Interaction experience; Interaction with avatars Simulation sickness; Side effects; Discomfort; Physical comfort of the equipment	Asghar et al. (2019), Barré et al. (2019), Beh et al. (2021), Benvegnù et al. (2021), Bhagwat et al. (2021), Borsci et al. (2016), Bunes and Karlens (2019), Burigat and Chittaro (2016), Butt et al. (2018), Buttussi and Chittaro (2021), Casso et al. (2019), Cavalcanti et al. (2021), Ceci et al. (2021), Chae et al. (2021), Chan et al. (2021), Cooper et al. (2021), Dhalmahapatra et al. (2021), Di Vece et al. (2021), Dorozhkin et al. (2017), Chen et al. (2021), Liang et al. (2018), Liang et al. (2019a), Liang et al. (2020), Zhang (2017), Xu et al. (2019b), Jaskiewicz et al. (2020), Joshi et al. (2021), Jung and Kim (2017), Zhang et al. (2017), Keßler et al. (2020), Kim et al. (2020), Kwegyir-Afful and Kantola (2021), Li et al. (2017), Nykänen et al. (2020), Oh et al. (2020), Ouyang et al. (2018), Pham et al. (2018), Polivka et al. (2019), Prendinger et al. (2016), Chen et al. (2020), Rahouti et al. (2021), Saghafian et al. (2020), Sankaranarayanan et al. (2016), Shewaga et al. (2018), Song et al. (2021), Sportillo et al. (2018), Weber et al. (2020), Xiao et al. (2020), Xu and Zheng (2021), Zhao et al. (2016)
Perceived learning effectiveness	41 (30.15%)	Helpfulness of system; Perceived benefits; Behavior influence; Perceived learning gains; Perceived usefulness; Perceived utility; Skills acquisition; Great tool for learning; Ease of learning; Comprehension; Educational value; Instruction efficacy; Easiness; Content relevancy and applicability (suitability); Applicability; Relevancy; Effectiveness; Efficiency; Cognitive benefits; Critical thinking	Ahn et al. (2020), Avveduto et al. (2017), Benda et al. (2020), Benvegnù et al. (2021), Bhagwat et al. (2021), Burigat and Chittaro (2016), Butt et al. (2018), Casso et al. (2019), Chae et al. (2021), Chan et al. (2021), Chang et al. (2019), Chittaro et al. (2018), Di Vece et al. (2021), Dorozhkin et al. (2017), Gupta and Varghese (2020), Chen et al. (2021), Liang et al. (2019a), Liang et al. (2020), Zhang (2017), Hoogenes et al. (2018), Jaskiewicz et al. (2020), Jung and Kim (2017), Keßler et al. (2020), Kwegyir-Afful and Kantola (2021), Mirauda et al. (2020), Nykänen et al. (2020), Ouyang et al. (2018), Pham et al. (2018), Polivka et al. (2019), Prasolova-Förlund et al. (2017), Prendinger et al. (2016), Chen et al. (2020), Rahouti et al. (2021), Wu et al. (2020b), Saghafian et al. (2020), Sankaranarayanan et al. (2016), Sim et al. (2019), Song et al. (2021), Sportillo et al. (2018), Truong et al. (2021), Vahdatikhaki et al. (2019), Zhao et al. (2016)
Subjective comparison	15 (11.03%)	Comparison with conventional training methods; Preference over other methods	Bhagwat et al. (2021), Bunes and Karlens (2019), Butt et al. (2018), Dorozhkin et al. (2017), Liang et al. (2019a), Liang et al. (2020), Zhang (2017), Jaskiewicz et al. (2020), Kwegyir-Afful and Kantola (2021), Lovreglio et al. (2021), Mondragón-Bernal (2020), Rahouti et al. (2021), Saghafian et al. (2020), Sankaranarayanan et al. (2016), Vahdatikhaki et al. (2019)

Table 1 (continued)

Categorized measure	No. of studies (%)	Raw measures	Studies
Presence	22.5 (16.54%)	Presence; Immersion*; Vividness; Experientiality	Avveduto et al. (2017), Barré et al. (2019), Benvegnù et al. (2021), Buttussi and Chittaro (2018), Buttussi and Chittaro (2021), Clifford et al. (2019), Cooper et al. (2021), Dhalmahapatra et al. (2021), Eiris et al. (2020), Fu and Li (2020), Zhang (2017), Jeelani et al. (2017), Joshi et al. (2021), Kim et al. (2020), Kwon (2020), Leder et al. (2019), Nykänen et al. (2020), Ouyang et al. (2018), Chen et al. (2020), Rahouti et al. (2021), Sharma et al. (2020), Song et al. (2021), Weber et al. (2020)
Engagement or interactive-ness	18 (13.24%)	Flow; Engagement; Involvement; Attention; Interactivity; Active learning; Sense of control; Immediacy of control; Awareness; Fixation	Ahn et al. (2020), Avveduto et al. (2017), Benvegnù et al. (2021), Bunes and Karlsen (2019), Burigat and Chittaro (2016), Butt et al. (2018), Buttussi and Chittaro (2018), Buttussi and Chittaro (2021), Cavalcanti et al. (2021), Chittaro et al. (2018), Comu et al. (2021), Zhang (2017), Kwon (2020), Mirauda et al. (2020), Pham et al. (2018), Chen et al. (2020), Saghafian et al. (2020), Salcedo et al. (2016), Shewaga et al. (2018)
Intention to use	20 (14.71%)	Possibility to recommend to others; Recommendation of VR; Attitude towards integrating VR; Intention to use; Innovativeness and adaptability to VR; Possibility to use as supplementary training; Likelihood to practice	Benda et al. (2020), Bunes and Karlsen (2019), Butt et al. (2018), Chae et al. (2021), Chan et al. (2021), Di Vece et al. (2021), Dorozhkin et al. (2017), Gupta and Varghese (2020), Chen et al. (2021), Liang et al. (2019a), Liang et al. (2020), Zhang (2017), Mondragon-Bernal (2020), Ouyang et al. (2018), Prasolova-Förland et al. (2017), Saghafian et al. (2020), Sankaranarayanan et al. (2016), Sim et al. (2019), Sportillo et al. (2018), Xiao et al. (2020)
Confidence level	19 (13.97%)	Confidence level; Self-efficacy; Enhanced readiness; Confidence building; Safety self-efficacy; Safety-related outcome expectations	Buttussi and Chittaro (2018), Buttussi and Chittaro (2021), Chittaro et al. (2018), Edwards et al. (2021), Gupta and Varghese (2020), Herrington and Tacy (2020), Lovreglio et al. (2021), Makransky et al. (2019), Nykänen et al. (2020), Rahouti et al. (2021), Wu et al. (2020b), Schulz et al. (2019), Song et al. (2021), Ta et al. (2019), Tang et al. (2021), Truong et al. (2021), Yau et al. (2021), Yu et al. (2021), Zhao et al. (2016)
Comments, impressions or feedbacks	23.5 (17.28%)	Overall experience and opinions; Impressions; First impressions; Advantages; Limitations; Recommendation for improvement; Recommendation for change; User comments; Feedback	Benda et al. (2020), Benvegnù et al. (2021), Bhagwat et al. (2021), Bunes and Karlsen (2019), Butt et al. (2018), Buttussi and Chittaro (2018), Buttussi and Chittaro (2021), Casso et al. (2019), Cavalcanti et al. (2021), Clifford et al. (2019), Di Vece et al. (2021), Din and Gibson (2019), Dorozhkin et al. (2017), Chen et al. (2021), Zhang (2017), Herrington and Tacy (2020), Jeelani et al. (2020), Nykänen et al. (2020), Prasolova-Förland et al. (2017), Chen et al. (2020), Saghafian et al. (2020), Shewaga et al. (2018), Shi et al. (2019), Shi et al. (2020)

*While the term “immersion” is used in some of the papers, the measure is typically referred to as presence

confidence level, and comments, impressions or feedback. These categories were developed by combining similar measures together, however, it is important to note that each category may not perfectly align with each included measure as each measure has its distinctive definition. Classifying each measure into the most appropriate category was deemed necessary to extract meaningful insights that aligned with Kirkpatrick's model. Researchers intending to perform an evaluation study using one of the measure categories are encouraged to assess the suitability of the exact measure to be used.

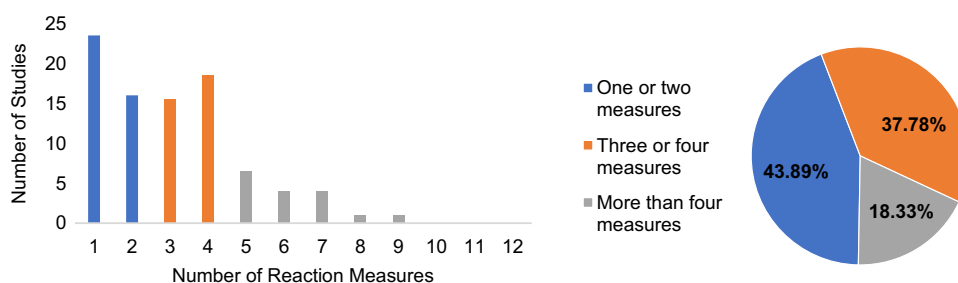
Out of the 136 studies reviewed, 90 (66.18%) evaluated at least one measure of reaction. Usability is the most common reaction measure used by 35.66% of the reviewed studies, followed by perceived learning effectiveness (30.15%) and affective reaction (20.59%) measures. Usability can be described as the “quality of a user's experience when interacting with products or systems” (U. S. General Services Administration n.d.a). Existing surveys such as the system usability scale (SUS Brooke 1996) and simulator sickness questionnaire (Kennedy et al. 1993) were commonly used, although other studies also measure other metrics such as ease of use and user-friendliness. Perceived learning effectiveness is the trainees' perception of how effective the VR training system is for learning. Metrics, such as the helpfulness of the system for learning, perceived benefits, learning gains, and usefulness, are included as part of this measure. Affective or emotional reaction comprises trainees' overall feeling toward the training, their perceived enjoyment, how pleasant, attractive, and impressive the training is, and their degree of satisfaction. In addition to these three measures, evaluators are interested in the general comments, impressions, and feedback from trainees in 17.28% of the studies. These measures are typically obtained by using interviews or open-ended post-surveys. This process allows trainees to express their perceived strengths and weaknesses of the system and recommend changes and improvements.

Presence represents the trainees' sense of being in a virtual environment, realism measures how realistic or similar a virtual environment is with respect to the real-world equivalent. The two measures were used for evaluation in 16.54% and 15.81% of all studies, respectively. Existing surveys, such as the Igroup Presence Questionnaire (IPQ Schubert

et al. 2001), Presence Questionnaire (PQ Witmer and Singer 1998), and the Slater–Usoh–Steed presence questionnaire (Usoh et al. 2000) are commonly used for evaluating presence. Intention to use, measures the desire for trainees to undertake or integrate the evaluated VR training in the future, their likelihood to practice using the VR system, and their likelihood to recommend the training to others. This measure is used in 14.71% of all studies. Confidence level represents 13.97% of the studies, this includes studies that measure “self-efficacy”, which is defined as one's belief in their own ability to execute the necessary behaviors to achieve specific goals (Bandura 1977). Engagement or interactivity includes metrics, such as flow, involvement, attention, active learning, and sense and immediacy of control. The measure of engagement is used in 13.24% of studies. Subjective comparison is measured by 11.03% of studies, this is defined as a trainee's subjective preference when comparing VR training with traditional training methods or other methods of interest. The final two measures, quality of instruction and feedback, and motivation, are used as evaluation measures in 8.82% and 5.88% of the studies, respectively.

Figure 9 illustrates a diminishing number of studies that evaluated a higher number of reaction measures with 43.89% of studies evaluating reaction using two or fewer reaction measures and 81.67% of studies using four or fewer measures. Usability and presence were the two leading measures for studies with two or fewer measures. Out of the 23.5 studies with only one measure, 25.53% measured usability and 21.28% measured presence. Out of the 16 studies with two measures, 28.13% measured usability and 21.88% measured presence. Usability and presence can provide validation that the VR system used was of an acceptable standard and are therefore recommended to be used in all studies evaluating the effectiveness of VR training. If a standardized questionnaire is used such as SUS for usability or PQ, IPQ, or Slater–Usoh–Steed presence questionnaire for presence then comparisons of usability and presence can be conducted between studies. However, these two measures, while important, are not very useful when measured in isolation when investigating training effectiveness. Rather, they should be complemented with other measures depending on the intended outcome of the training. For example,

Fig. 9 Number of studies measuring reaction and the number of reaction measures evaluated (out of 90 studies that measured reaction)



if training is intended to increase trainees' confidence in performing a task, then confidence level questions should be included in addition to usability and presence. Similarly, if training is intended to promote trainees' intention to train using the system or to increase engagement during training, then the respective measures should be implemented. In addition to providing a subjective assessment of training, evaluating reaction has the added benefit of identifying areas of the training that need improvement. For example, if the realism of a VR training system is rated inadequately, then realism should be the focus for future improvement. This benefit can be capitalized by evaluating a comprehensive set of measures rather than focusing on a single measure, such as usability or presence. Additionally, greater insights can be obtained by providing open-ended sections, inviting trainees to elaborate on their decisions.

4.4.2 Level 2—learning

Evaluations that measured learning used three methods—these include using pen-and-paper or computer-based knowledge tests, evaluating trainees' performance in the VR system, and measuring the performance in more realistic settings (i.e., transfer test). As previously shown in Fig. 8, 98 (72.06%) of all studies reviewed evaluated learning using either one or a combination of these methods. Performance test in VR was the most common method, used in 58 (42.65%) of all studies. This was followed by studies that measured knowledge with 33.5 (24.63%) of all reviewed studies utilizing a knowledge test. Finally, a transfer test was implemented by 26 (19.12%) of the studies (Fig. 10). Further analyses of each of the three learning methods are presented in the following order: knowledge test, performance test in VR, and transfer test.

Studies that evaluated learning using a knowledge test after training (post-survey) represented 24.63% of the 136 studies reviewed. In addition to measuring knowledge

post-training, several studies performed a test prior to training using a pre-survey with some also including a knowledge test at an extended period after training (e.g., one month) using a follow-up survey. A pre-survey knowledge test is useful in measuring an individual's baseline knowledge prior to training, this helps determine the different levels of knowledge between trainees and knowledge gained from the training being evaluated. This test is recommended when evaluation includes training subjects who are likely to have prior knowledge or experience directly related to the training (Kirkpatrick and Kirkpatrick 2007, p. 49). A baseline measurement for the outcome variable, in this case, the knowledge level is always recommended in general to account for any unequal distribution even in a randomized true-experimental design (Shannon et al. 1999). As observed in Table 2, out of 33.5 studies that performed a post-knowledge test, almost half (43.28%) of them did not obtain baseline knowledge. A follow-up survey administering a knowledge test is useful to gain insights into how well the knowledge is retained after an extended period. This is particularly important for safety training where refresher training is recommended as retention results are useful in determining training frequency and period.

A total of 58 (42.65%) of all studies reviewed evaluated learning by measuring trainee performance in VR either during or after training. Performance measures are domain-specific in nature as VR training is used in different application domains (Sect. 4.1). However, the measures can be generalized into categories such as economy or efficiency, general performance, safety, completion or pass rate, autonomy, response time, and hazard identification and management. Similar to reaction measures, each category is defined by combining similar (but slightly different) measures together for the purpose of this review. A list of the measures classified into the categories is available in Table 3, which is also applicable to the transfer test described later in this section. Table 4 presents the studies using each categorized measure for both the performance test in VR and the transfer test.

General performance or accuracy measured using a score was the most commonly used measure when

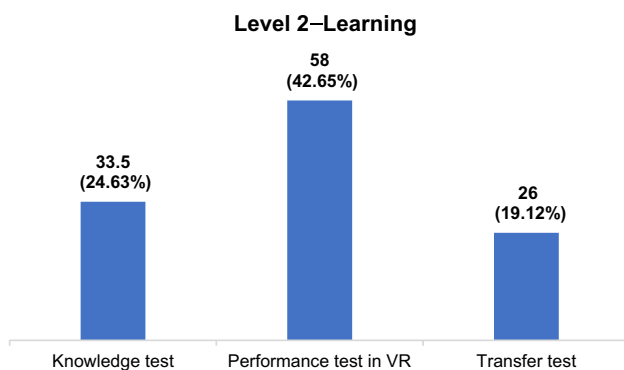


Fig. 10 Learning methods used by review studies as a percentage of use

Table 2 Number and percentage of studies that evaluate learning using knowledge test before, directly after, and an extended period after training

Time point	No. of studies (%)
Pre	20 (14.71%)
Post	33.5 (24.63%)
Follow-up	8.5 (6.25%)
Pre and post	19 (13.97%)
Post and follow-up	8.5 (6.25%)
Pre, post, and follow-up	4.5 (3.31%)

Table 3 Categorized performance test in VR and transfer test measures

Categorized measure	Raw measures
Economy or efficiency	Completion time (speed); Economy; Efficiency of movement; Efficiency of usage (wastage); Number of attempts; Number of repetitions until a predetermined proficiency is achieved; Time spent training
General performance or accuracy	Total performance score; Domain-specific performance graded by experts; Accuracy; Error rate; Negative accuracy
Safety	Safety; Hazard exposure time; Time to risk evasion; Physical damage score; Safety negative performance score
Completion or pass rate	Completion rate; Pass rate; Success rate; Failure rate
Autonomy	Autonomy (ability to perform task without instructions, hints, or help)
Response time	Response/reaction time
Hazard identification or management	Hazard identification index (e.g., number of hazards identified); Hazard management; Fault diagnosis
Others	Foot placement, clearance, and peak elevation; Walking patterns, speed trend, and time; Risk-taking decision-making; Postural stability

evaluating performance in VR and was used by 39 (28.68%) of all studies. The number of errors during training is also another key metric used to measure performance. Studies that evaluated a multivariate score for total performance, whether it was calculated systematically within the VR system or graded by experts, was also a commonly used measure of performance.

Economy or efficiency was used to evaluate learning by 31 (22.79%) of all studies reviewed and covered the efficiency of time, movement, resource usage, and the number of attempts until achieving a predetermined proficiency. Efficiency of time is measured as the time spent during training or the completion time to perform an assessment task. Trainees are deemed to be proficient if they can perform tasks accurately and quickly. The efficiency of movement is evaluated within a VR system by calculating the number of movements and the total path distance of these movements. This measure is relevant for training psychomotor skills where precise movements are necessary to maintain safety (e.g., surgical procedures). Reduced resource usage, wastage and the number of attempts are also important measures used to evaluate economy and efficiency.

A safety score, including negative scores for physical damage (in the virtual environment) to the trainees and others (e.g., virtual patients), was used by 9.56% of all reviewed studies. Time spent exposed to virtual hazards and the time used to avoid them were also common measures of safety.

Some performance tasks involved using a predetermined benchmark measure for which a pass or fail was determined. This measure allowed trainees to complete the task and compare the final score with a predetermined benchmark (i.e., pass or fail). Other studies terminated the task once mistakes are made (i.e., complete, or incomplete). Studies that evaluated the effect of VR training using completion and pass rates accounted for 6.62%. Similarly, studies that evaluated the ability of trainees to identify, diagnose, and manage virtual hazards accounted for 6.62% of the studies.

A small portion of the studies reviewed evaluated the response or reaction time of the trainees (2.21%) and their autonomy or the ability to perform the task without additional help (1.47%).

The fact that performance tests conducted in VR were the most used method for evaluating learning suggests another opportunity for VR for training assessments. VR offers not only the ability to deliver training, but also to objectively assess competency. The ability to simulate reality safely is beneficial for both simulating realistic hands-on training and assessment. The prime example of simulation for assessment is in the aviation industry where prospective pilots are required to log their time on the simulators to be qualified (Bradley and Abelson 1995). Rapid advancements in VR (second VR wave) have brought a range of new VR technologies leading to greater interest in a more diverse set of fields, future studies reviewing the current state of VR as an assessment tool for competency may provide valuable insights.

A transfer test is a performance test conducted in a similar environment to its comparative real task. This measure can be administered by asking trainees to perform the actual task or a simulated version of the task. In the health services domain, a simulation may be performed on a manikin, cadaver, 3D-printed or artificial physical model while the actual task may be performed via surgery or assisting a supervised procedure. As presented in Fig. 10, 26 (19.12%) of all studies reviewed evaluated transfer. Similar to the performance test in VR, general performance or accuracy measured using a score was the most common measure used in 15.44% of all studies, followed by economy and efficiency in 8.46% of studies. With regards to evaluations of economy and efficiency using a transfer test, all included studies measured the task completion time. Domain-specific performance graded by experts was commonly used for performance scores, particularly for health services-related training where standardized assessment rating scales exist. These scales are

Table 4 Studies categorized by performance test in VR (PVR) and transfer test (TRF)

Categorized measure	Evaluation method	No. of studies (%)	Studies
Economy or efficiency	PVR	31 (28.68%)	Wang et al. (2021), Edwards et al. (2021), Beh et al. (2021), Xiao et al. (2020), Wu et al. (2020b), Shi et al. (2020), Fu and Li (2020), Casso et al. (2019), Brown et al. (2019), Yang et al. (2018), Shim et al. (2018), Shewaga et al. (2018), Sebastian et al. (2018), Nilsson et al. (2017), Waterman et al. (2016), Colombo and Golzio (2016), Burigat and Chittaro (2016), Di Vece et al. (2021), Chen et al. (2021), Fu et al. (2017), Francone et al. (2019), Caporusso et al. (2019), Liu et al. (2016), Bing et al. (2019), Xu et al. (2019b), Neis et al. (2016), Borsci et al. (2016), Schulz et al. (2019), Sinitzky et al. (2020), Truong et al. (2021), Smith and Veitch (2019)
	TRF	11.5 (8.46%)	Yau et al. (2021), Xiao et al. (2020), Hoogenes et al. (2018), Orland et al. (2020), Waterman et al. (2016), Cooper et al. (2021), Shim et al. (2018), Chittaro et al. (2018), Jensen et al. (2016), Mirauda et al. (2020), Logishetty et al. (2019), Xin et al. (2019)
General performance or accuracy	PVR	39 (28.68%)	Beh et al. (2021), Borsci et al. (2016), Brown et al. (2019), Burigat and Chittaro (2016), Caporusso et al. (2019), Cavalcanti et al. (2021), Di Vece et al. (2021), Edwards et al. (2021), Francone et al. (2019), Fu and Li (2020), Gawecki et al. (2020), Gibson et al. (2020), Liang et al. (2018), Wu et al. (2020a), Havemann et al. (2019), Xu et al. (2019b), Wang et al. (2021), Lang et al. (2018), Liu et al. (2016), Neis et al. (2016), Fu et al. (2017), Wu et al. (2020b), Salcedo et al. (2016), Schulz et al. (2019), Sebastian et al. (2018), Simpson and Rafferty (2020), Sinitzky et al. (2020), Smith and Veitch (2019), Smith et al. (2019), Sportillo et al. (2018), Sugand et al. (2016), Suto et al. (2020), Tawadrous et al. (2017), Truong et al. (2021), Tsuboi et al. (2018), Xiao et al. (2020), Shi et al. (2020), Wang et al. (2018)
	TRF	21 (15.44%)	Yau et al. (2021), Xiao et al. (2020), Hauschild et al. (2021), Logishetty et al. (2019), Frendo et al. (2021), Hoogenes et al. (2018), Gawecki et al. (2020), Waterman et al. (2016), Edwards et al. (2021), Rossler et al. (2019), García et al. (2016), Makransky et al. (2019), Adami et al. (2021), Zhang et al. (2017), Jensen et al. (2016), Jaskiewicz et al. (2020), Xin et al. (2019), Jung and Ahn (2018), Mirauda et al. (2020), Orland et al. (2020), Cooper et al. (2021), Chittaro et al. (2018)
Safety	PVR	13 (9.56%)	Brown et al. (2019), Cavalcanti et al. (2021), Colombo and Golzio (2016), Francone et al. (2019), Wang et al. (2021), Li et al. (2017), Neis et al. (2016), Fu et al. (2017), Schulz et al. (2019), Sportillo et al. (2018), Tsuboi et al. (2018), Wang et al. (2018), Yang et al. (2018)
	TRF	1 (0.74%)	Jensen et al. (2016)
Completion or pass rate	PVR	9 (6.62%)	Beh et al. (2021), Borsci et al. (2016), Brown et al. (2019), Fu and Li (2020), Havemann et al. (2019), Wang et al. (2021), LoJacono et al. (2018), Neis et al. (2016), Truong et al. (2021)
	TRF	5 (3.68%)	Orland et al. (2020), Yau et al. (2021), Gawecki et al. (2020), Xin et al. (2019), Xin et al. (2020)
Autonomy	PVR	2 (1.47%)	Edwards et al. (2021), Colombo and Golzio (2016)
	TRF	1 (0.74%)	Orland et al. (2020)
Response time	PVR	3 (2.21%)	Cavalcanti et al. (2021), Lang et al. (2018), Sportillo et al. (2018)
	TRF	0 (0%)	
Hazard identification or management	PVR	9 (6.62%)	Agrawal et al. (2017), Colombo and Golzio (2016), Dhalmahapatra et al. (2021), Eiris et al. (2020), Gupta and Varghese (2020), Jeelani et al. (2020), Li et al. (2017), Moore et al. (2019), Polivka et al. (2019)
	TRF	1 (0.74%)	Leder et al. (2019)
Others	PVR	2 (1.47%)	LoJacono et al. (2018), Shi et al. (2019)
	TRF	3 (2.21%)	Leder et al. (2019), LoJacono et al. (2018), Cyma-Wejchenig et al. (2020)

used as part of a traditional training assessment where the assessor observes each trainee's performance and provides a score for each of the competencies listed. Examples of such scales include the Arthroscopic Surgical Skill Evaluation Tool and Cochlear Implant Surgery Assessment Tool. Five studies measured the completion or pass rate, and another three studies measured either safety, autonomy or hazard identification and management.

4.4.3 Level 3—behavior

None of the included studies evaluated behavior. The lack of studies evaluating this category (Level 3) is reasonable and likely due to the logistical requirement for a specialized trainer to observe the job behavior of multiple trainees.

4.4.4 Level 4—results

Three studies, namely, García et al. (2016), Wu et al. (2020b), and Butt et al. (2018) performed result (level 4) evaluations. García et al. (2016) evaluated the reported number of accidents, working days lost, and reduction in expenses after the implementation of a desktop-based VR training experience for the maintenance and operations of high-voltage overhead power lines. The data was gathered between one and three years after the implementation of the VR training. Wu et al. (2020b) recorded the number of self-reported injuries two months after undertaking a needle stick or sharp injury prevention VR training experience for nursing and medical students. Butt et al. (2018) focused on whether the use of VR for urinary catheterization training would encourage practice and recorded the time each trainee spent practicing after training. However, the considered time interval for recording the result was short (i.e., two-week intervals).

5 General discussions, recommendations, and limitations

This review found that existing studies that evaluated the effectiveness of VR for safety-relevant training were mostly applicable to health services and construction domains. Although health services and construction domains represented approximately 55% of the studies reviewed, interests from a wide range of domains suggest that VR may provide a wide range of benefits to several fields. Surgical training, an early adopter of VR for training, had the greatest number of effectiveness evaluation studies. The majority of these studies also utilized either true or quasi-experimental design, which is considered to provide more reliable results than non-experimental design. Health services, particularly surgery, also had the greatest number of VR applications for safety-relevant training. In addition to the availability and

applicability of commercially available surgery simulators, the dangerous nature of surgical procedures is likely to be an important contributing factor that encourages researchers in this domain to perform rigorous evaluations and validations of the system. Similarly in construction, the inherent dangers to people within the industry as indicated by high fatality rates have likely led to an appetite for innovation in safety training delivery such as exploring the benefits of using VR. While interests have been strong within the construction domain, the industry lacks a standardized approach to its use of VR in training and applicable simulators, with the majority of studies performing evaluation using a non-experimental design for one-off prototypes. The difference in evaluation design between health services and construction domains may be due to both domains having different levels of experience in implementing virtual training simulators, particularly from a historical point of view. Researchers investigating the construction domain are encouraged to perform evaluations that implement true-experimental design to improve the reliability of results. If effectiveness is validated using a true-experimental design, it may motivate others working in the construction sector to invest in VR technology and may encourage an effort towards standardization and encourage the development of commercially available construction safety simulators.

Despite being the domain with the highest fatality rate both in the USA and in Australia, no study was found in the agriculture, fishing, and forestry sector. The ineffectiveness of current traditional training methods for preventing accidents in this sector should prompt researchers to investigate alternative solutions such as investigating the benefits of using VR technology for training. The high number of self-employed workers in this sector, also with limited access to safety officers and training facilities can potentially leverage VR training that is system-based, less reliant on instructors, and flexible in terms of the time and place of training. Developing VR training on how to safely operate vehicles such as tractors should not be difficult considering there are other vehicle simulators being developed for different domains. As with other domains, the challenge is in the successful implementation of training systems such that workers in the sector improve their safe behavior, which eventually results in fewer accidents and fatalities. Future research should aim to investigate the implementation of VR safety training in the agriculture, fishing, and forestry sectors to evaluate whether it can provide benefits towards safety-related training.

The majority of studies reviewed focus on evaluating the effectiveness of VR training as opposed to prototype development. This result may suggest that VR is starting to mature into an effective training tool and outgrowing its infancy stage. Despite this, a significant portion of the studies reviewed were still focused on evaluating the general usability of the hardware and side effects of the technology,

especially for studies that focused on prototyped systems. Many studies concentrated on evaluating the effects of specific features of VR and their effect on training effectiveness rather than the general effectiveness of the training system. This is likely due to the technology still being relatively new with many questions being asked between the interplay of different features such that each feature is best evaluated individually. As VR matures and becomes more reliable, each of these features is likely to be better understood, allowing evaluators to focus more on the effectiveness of the system in delivering training. The combination of studies focusing on development and evaluation is crucial to push VR technology forward. A developmental study focuses on testing new technology whereas an evaluation study either provides validation or rejects that the technology being investigated is effective and that efforts in further advancements are worthwhile. While effectiveness evaluation is beyond the scope of some developmental studies, it is recommended that when evaluation is performed, post-only one-group design is avoided by taking additional steps to satisfy quasi-experimental design requirements.

This review categorized evaluation measures using Kirkpatrick's four-level model to determine how studies currently evaluate the effectiveness of VR safety-relevant training, categories included reaction, learning, behavior, and result measures. The majority of studies evaluated the effectiveness of VR safety-relevant training in terms of its learning (level 2 72.06%) followed closely by reaction (level 1 66.18%). There were no studies reviewed that evaluated the effectiveness using behavior measures (level 3) with a limited number of studies evaluating using result measures (level 4). The limited number of studies evaluating effectiveness using behavior and result measures is not surprising considering they are substantially more complicated, time-consuming, and resource intensive to implement these measures compared to reaction and learning measures. In fact, Kirkpatrick considered evaluating behavior to be "the most difficult and time consuming of the four levels" (Kirkpatrick and Kirkpatrick 2007, p. 105). The categories for outcome measures are presented both in Sect. 4.4.1. (Level 1–reaction) and Sect. 4.4.2. (Level 2–learning) and aim to provide some guidance for researchers and assist in making informed choices about evaluating training effectiveness. When reaction is measured, it is recommended to implement designs that evaluate the usability of the VR system and the sense of presence using established questionnaires. This approach is likely to provide readers with confidence that the evaluation of the VR training has been performed on a VR system of an acceptable standard. Additionally, when established questionnaires are used appropriately, researchers can estimate the quality of their VR training by comparing the values of these results against existing systems. If the intended outcome of using VR technology in training is to improve learning outcomes, then it is recommended to use at least one

of the three commonly used evaluation methods which are knowledge test, performance test in VR, and transfer test. It is important that each training solution is appropriately evaluated using a test that best represents the objective of the training. For example, if an assembly training solution aims to teach trainees how to perform real assembly tasks, then performing a knowledge test or a performance test in VR may not be sufficient, instead, it is likely a transfer test is required to test their ability after training. On the other hand, if a training solution aims to convey important information, then a knowledge test should suffice. Finally, while the common aim for most safety training is to reduce the number of accidents and fatalities, which requires an evaluation of results (level 4), it is reasonable that there is still a lack of studies evaluating training using results measures due to the difficult nature in its implementation particular with a relatively new research area. Efforts toward evaluating the implementation of VR safety-relevant training in the workplace by observing trainees' behavior (level 3) and the results from implementation (level 4) should be a focus of future research.

One of the limitations of this review is the potential overlook of relevant studies due to the search strategy. Although this paper attempts to comprehensively identify all relevant studies, applying restrictions in the scope of the search using a combination of search strings is needed to ensure the feasibility of this review. As a result, some relevant studies may possibly be unidentified. For example, this review includes studies that use the form of the word "safe" and "train" to generally cover safety training. Studies that focus on safety-relevant training such as disaster preparedness training, or other types of professional skills training may not recognize or explicitly state the safety implications of their training and thus, not include the term "safe" in their paper. Similarly, studies that use different terminology for their training outcomes instead of "effectiveness", "impact", or "outcome" may also be overlooked. Another limitation is the exclusion of studies prior to 2016, which may limit the generalizability of the findings, specifically to VR safety-relevant training prior to 2016. Finally, the current review does not analyze the findings and assess the quality of the studies (beyond study design). Investigating whether VR for safety-relevant training is effective as well as assessing the quality of the evaluation processes and the findings may provide valuable insights. These are beyond the scope of the current review and present an important opportunity for future research to address.

6 Conclusion

This review presents the current state of evaluation studies that focus on the effectiveness of VR safety-relevant training. Research domains in which VR safety-relevant training was applied are identified and described with results showing

that the health services domain followed by construction is the most active domain. The study objective and design of each evaluation study are analyzed with results showing that the majority of studies focus on evaluation, which utilized true- or quasi-experimental design. Besides performing evaluation studies, developmental studies should also be performed to continue to push the boundaries of what VR technology can provide, however, it is recommended that these studies also perform evaluation of their prototypes using an acceptable experimental design (e.g., true- or quasi-experimental design). Finally, the paper categorizes evaluation measures for each study using the four levels of the widely used Kirkpatrick's model. Reaction (level 1) and learning (level 2) are the two most commonly used evaluation measures while none of the reviewed studies evaluated behavior (level 3) and only three studies evaluated results (level 4). Usability, perceived learning effectiveness, and affective reactions are the three most commonly used reaction (level 1) measures. When evaluating learning (level 2), there are three commonly used evaluation methods, these include knowledge test, performance test in VR, and transfer test. Categories of measures are also identified and analyzed for the performance test in VR and the transfer test. By describing the space of VR safety-relevant evaluation studies, this review aims to extend the existing body of literature in providing information and guidance on how such VR training evaluations can be performed. Discussions and recommendations for future work are presented with the aim to achieve more comprehensive, standardized, and consistent evaluations when measuring the effectiveness of VR for safety-related training.

Author contributions HS: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization. MM: Conceptualization, Validation, Writing—Original Draft, Writing—Review & Editing, Supervision, Project administration. BH: Conceptualization, Supervision, Project administration, Funding acquisition.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest This study was co-funded by the Department of Industry, Science, Energy and Resources (Innovative Manufacturing CRC Ltd) and Melbourne Water Corporation (IMCRC/MWC/131020). KGSupport reviewed the article and provided editor support services.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source,

provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abich J, Parker J, Murphy JS, Eudy M (2021) A review of the evidence for training effectiveness with virtual reality technology. *Virtual Real.* <https://doi.org/10.1007/s10055-020-00498-8>
- Adami P, Rodrigues PB, Woods PJ, Becerik-Gerber B, Soibelman L, Copur-Gencturk Y, Lucas G (2021) Effectiveness of vr-based training on improving construction workers' knowledge, skills, and safety behavior in robotic teleoperation. *Adv Eng Inform* 50:101431. <https://doi.org/10.1016/j.aei.2021.101431>
- Agrawal R, Knodler M, Fisher DL, Samuel S (2017) Advanced virtual reality based training to improve young drivers' latent hazard anticipation ability. *Proc Human Factors Ergon Soc Annu Meet* 61(1):1995–1999. <https://doi.org/10.1177/1541931213601994>
- Ahn S, Kim T, Park YJ, Kim JM (2020) Improving effectiveness of safety training at construction worksite using 3d bim simulation. *Adv Civil Eng* 2020:2473138. <https://doi.org/10.1155/2020/2473138>
- Aïm F, Lonjon G, Hannouche D, Nizard R (2016) Effectiveness of virtual reality training in orthopaedic surgery. *Arthroscopy* 32(1):224–232. <https://doi.org/10.1016/j.arthro.2015.07.023>
- Alaker M, Wynn GR, Arulampalam T (2016) Virtual reality training in laparoscopic surgery: a systematic review & meta-analysis. *Int J Surg* 29:85–94. <https://doi.org/10.1016/j.ijso.2016.03.034>
- Alliger GM, Janak EA (1989) Kirkpatrick's level of training criteria: Thirty years later. *Pers Psychol* 42(2):331–342. <https://doi.org/10.1111/j.1744-6570.1989.tb00661.x>
- Alvarez K, Salas E, Garofano CM (2004) An integrated model of training evaluation and effectiveness. *Hum Resour Dev Rev* 3(4):385–416. <https://doi.org/10.1177/1534484304270820>
- Asghar I, Egaji OA, Dando L, Griffiths M, Jenkins P (2019) A virtual reality based gas assessment application for training gas engineers. In: Proceedings of the 9th international conference on information communication and management (pp 57–61) Association for Computing Machinery. <https://doi.org/10.1145/3357419.3357443>
- Avveduto G, Tanca C, Lorenzini C, Tecchia F, Carrozzino M, Bergamasco M (2017) Safety training using virtual reality: a comparative approach. In: DePaolis LT, Bourdot P, Mongelli A (ed), *Avr 2017: International conference on augmented reality, virtual reality and computer graphics* pp 10324:148–163 Springer Nature. https://doi.org/10.1007/978-3-319-60922-5_11
- Bandura A (1977) Self-efficacy: Toward a unifying theory of behavioral change. *Psychol Rev* 84(2):191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Barré J, Michelet D, Truchot J, Jolivet E, Recanzone T, Stiti S, Tesnière A, Pourcher G (2019) Virtual reality single-port sleeve gastrectomy training decreases physical and mental workload in novice surgeons: an exploratory study. *Obes Surg* 29(4):1309–1316. <https://doi.org/10.1007/s11695-018-03680-9>
- Bates R (2004) A critical analysis of evaluation practice: the kirkpatrick model and the principle of beneficence. *Eval Program*

- Plann 27(3):341–347. <https://doi.org/10.1016/j.evalprogplan.2004.04.011>
- Beh HJ, Rashidi A, Talei A, Lee YS (2021) Developing engineering students' capabilities through game-based virtual reality technology for building utility inspection. *Eng Constr Archit Manag.* <https://doi.org/10.1108/ecam-02-2021-0174>
- Benda NC, Kellogg KM, Hoffman DJ, Fairbanks RJ, Auguste T (2020) Lessons learned from an evaluation of serious gaming as an alternative to mannequin-based simulation technology: Randomized controlled trial. *JMIR Ser Games.* 8(3):e21123. <https://doi.org/10.2196/21123>
- Benvengù G, Furlan M, Orso V, Gamberini L (2021) The role of virtual spaces and interactivity in emergency training. In: European conference on cognitive ergonomics 2021 (pp 1–5) Association for Computing Machinery. <https://doi.org/10.1145/3452853.3452881>
- Bhagwat K, Kumar P, Delhi VS (2021) Usability of visualization platform-based safety training and assessment modules for engineering students and construction professionals. *J Civil Eng Educ.* 147(2):04020016. [https://doi.org/10.1061/\(ASCE\)EI.2643-9115.0000034](https://doi.org/10.1061/(ASCE)EI.2643-9115.0000034)
- Bing EG, Parham GP, Cuevas A, Fisher B, Skinner J, Mwanahamuntu M, Sullivan R (2019) Using low-cost virtual reality simulation to build surgical capacity for cervical cancer treatment. *J Glob Oncol* 5:1–7. <https://doi.org/10.1200/JGO.18.00263>
- Blake M (1996) The nasa advanced concepts flight simulator-a unique transport aircraft research environment. In: AIAA Flight simulation technologies conference, San Diego, CA, USA, 385–392. <https://doi.org/10.2514/6.1996-3518>
- Borsci S, Lawson G, Jha B, Burges M, Salantri D (2016) Effectiveness of a multidevice 3d virtual environment application to train car service maintenance procedures. *Virtual Real* 20(1):41–55. <https://doi.org/10.1007/s10055-015-0281-5>
- Bradley DR, Abelson SB (1995) Desktop flight simulators: Simulation fidelity and pilot performance. *Behav Res Methods Instrum Comput* 27(2):152–159. <https://doi.org/10.3758/BF03204721>
- Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T (2016) De-duplication of database search results for systematic reviews in endnote. *J Med Libr Assoc* 104(3):240–243. <https://doi.org/10.3163/1536-5050.104.3.014>
- Brooke J (1996) Sus: A quick and dirty usability scale. In Jordan PW, Thomas B, Weerdmeester BA, McClelland AL (ed) Usability evaluation in industry (pp 189–195) CRC Press. <https://doi.org/10.1201/9781498710411-35>
- Brown C, Robinson D, Egan R, Hopkins L, Abdelrahman T, Powell A, Pollitt MJ, Lewis WG (2019) Prospective cohort study of haptic virtual reality laparoscopic appendectomy learning curve trajectory. *J Laparoendosc Adv Surg Tech* 29(9):1128–1134. <https://doi.org/10.1089/lap.2019.0332>
- Bunæs TH, Karlsen J (2019) Using single player virtual simulations for training on collaborative medical practice. In: Proceedings of the european conference on games-based learning (pp 119–126) Academic Conferences International. <https://doi.org/10.34190/GBL.19.095>
- Bureau of Labor Statistics (2016) National census of fatal occupational injuries in 2015. https://www.bls.gov/news.release/archives/cfoi_12162016.pdf
- Bureau of Labor Statistics (2021) National census of fatal occupational injuries in 2020. <https://www.bls.gov/news.release/pdf/cfoi.pdf>
- Burigat S, Chittaro L (2016) Passive and active navigation of virtual environments vs. traditional printed evacuation maps: a comparative evaluation in the aviation domain. *Int J Hum Comput Stud* 87:92–105. <https://doi.org/10.1016/j.ijhcs.2015.11.004>
- Burke MJ, Sarpy SA, Smith-Crowe K, Chan-Serafin S, Salvador RO, Islam G (2006) Relative effectiveness of worker safety and health training methods. *Am J Public Health* 96(2):315–324. <https://doi.org/10.2105/AJPH.2004.059840>
- Burke MJ, Salvador RO, Smith-Crowe K, Chan-Serafin S, Smith A, Sonesh S (2011) The dread factor: How hazards and safety training influence learning and performance. *J Appl Psychol* 96(1):46–70. <https://doi.org/10.1037/a0021838>
- Butt AL, Kardong-Edgren S, Ellertson A (2018) Using game-based virtual reality with haptics for skill acquisition. *Clin Simul Nurs* 16:25–32. <https://doi.org/10.1016/j.ecns.2017.09.010>
- Buttussi F, Chittaro L (2018) Effects of different types of virtual reality display on presence and learning in a safety training scenario. *IEEE Trans Visual Comput Graphics* 24(2):1063–1076. <https://doi.org/10.1109/TVCG.2017.2653117>
- Buttussi F, Chittaro L (2021) A comparison of procedural safety training in three conditions: Virtual reality headset, smartphone, and printed materials. *IEEE Trans Learn Technol* 14(1):1–15. <https://doi.org/10.1109/TLT.2020.3033766>
- Caporusso N, Biasi L, Cinquepalmi G, Bevilacqua V (2019) An immersive environment for experiential training and remote control in hazardous industrial tasks. In: Advances in intelligent systems and computing (Vol 795, pp 88–97) Springer Nature. https://doi.org/10.1007/978-3-319-94619-1_9
- Casso G, Schoettker P, Savoldelli GL, Azzola A, Cassina T (2019) Development and initial evaluation of a novel, ultraportable, virtual reality bronchoscopy simulator: the computer airway simulation system. *Anesth Analg* 129(5):1258–1264. <https://doi.org/10.1213/ANE.0000000000003316>
- Cavalcanti J, Valls V, Contero M, Fonseca D (2021) Gamification and hazard communication in virtual reality: a qualitative study. *Sensors* 21(14):4663. <https://doi.org/10.3390/s21144663>
- Cecil J, Kauffman S, Gupta A, McKinney V, Pirela-Cruz MDM (2021) Design of a human centered computing (hcc) based virtual reality simulator to train first responders involved in the covid-19 pandemic. In: 2021 IEEE international systems conference (syscon) (pp 1–7) IEEE. <https://doi.org/10.1109/SysCon48628.2021.9447090>
- Chae CJ, Kim D, Lee HT (2021) A study on the analysis of the effects of passenger ship abandonment training using vr. *Appl Sci* 13(1–15):5919. <https://doi.org/10.3390/app11135919>
- Chan HY, Chang HC, Huang TW (2021) Virtual reality teaching in chemotherapy administration: randomised controlled trial. *J Clin Nurs* 30(13–14):1874–1883. <https://doi.org/10.1111/jocn.15701>
- Chang CY, Sung HY, Guo JL, Chang BY, Kuo FR (2019) Effects of spherical video-based virtual reality on nursing students' learning performance in childbirth education training. *Interact Learn Environ.* <https://doi.org/10.1080/10494820.2019.1661854>
- Chapman LJ, Husberg B (2008) Agriculture, forestry, and fishing sector. *J Saf Res* 39(2):171–173. <https://doi.org/10.1016/j.jsr.2008.02.008>
- Chen Q, Low SE, Yap JWE, Sim AKX, Tan YY, Kwok BWJ, Lee JSA, Tan CT, Loh WP, Loo BLW, Wong ACK (2020) Immersive virtual reality training of bioreactor operations. In: Proceedings of 2020 IEEE international conference on teaching, assessment, and learning for engineering (tale 2020) (pp 873–878) IEEE. <https://doi.org/10.1109/TALE48869.2020.9368468>
- Chen H, Hou L, Zhang GK, Moon S (2021) Development of bim, iot and ar/vr technologies for fire safety and upskilling. *Autom Constr* 125:103631. <https://doi.org/10.1016/j.autcon.2021.103631>
- Chittaro L, Corbett CL, McLean GA, Zangrando N (2018) Safety knowledge transfer through mobile virtual reality: a study of aviation life preserver donning. *Saf Sci* 102:159–168. <https://doi.org/10.1016/j.ssci.2017.10.012>
- Clifford RMS, Jung S, Hoerrmann S, Billinqhurst M, Lindeman RW (2019) Creating a stressful decision making environment for

- aerial firefighter training in virtual reality. In: 26th IEEE Conference on virtual reality and 3d user interfaces, vr 2019—proceedings (pp 181–189) IEEE. <https://doi.org/10.1109/VR.2019.8797889>
- Colombo S, Golzio L (2016) The plant simulator as viable means to prevent and manage risk through competencies management: Experiment results. *Saf Sci* 84:46–56. <https://doi.org/10.1016/j.ssci.2015.11.021>
- Comu S, Kazar G, Marwa Z (2021) Evaluating the attitudes of different trainee groups towards eye tracking enhanced safety training methods. *Adv Eng Inform* 49:101353. <https://doi.org/10.1016/j.aei.2021.101353>
- Cooper N, Millela F, Cant I, White MD, Meyer G (2021) Transfer of training—virtual reality training with augmented multisensory cues improves user experience during training and task performance in the real world. *Public Library Sci One* 16(3):1–22. <https://doi.org/10.1371/journal.pone.0248225>
- Cyma-Wejchenig M, Tarnas J, Marciniak K, Stemplewski R (2020) The influence of proprioceptive training with the use of virtual reality on postural stability of workers working at height. *Sensors* 20(13):3731. <https://doi.org/10.3390/s20133731>
- Czarnek G, Strojny P, Strojny A, Richter M (2020) Assessing engagement during rescue operation simulated in virtual reality: a psychophysiological study. *Int J Human Comput Interact* 36(5):464–476. <https://doi.org/10.1080/10447318.2019.1655905>
- Detmer FJ, Hettig J, Schindele D, Schostak M, Hansen C (2017) Virtual and augmented reality systems for renal interventions: A systematic review. *IEEE Rev Biomed Eng* 10:78–94. <https://doi.org/10.1109/rbme.2017.2749527>
- Dhalmahapatra K, Maiti J, Krishna OB (2021) Assessment of virtual reality based safety training simulator for electric overhead crane operations. *Saf Sci* 139:105241. <https://doi.org/10.1016/j.ssci.2021.105241>
- Di Vece C, Luciano C, De Momi E (2021) Psychomotor skills development for veress needle placement using a virtual reality and haptics-based simulator. *Int J Comput Assist Radiol Surg* 16(4):639–647. <https://doi.org/10.1007/s11548-021-02341-0>
- Din ZU, Gibson GE Jr (2019) Serious games for learning prevention through design concepts: an experimental study. *Saf Sci* 115:176–187. <https://doi.org/10.1016/j.ssci.2019.02.005>
- Dorozhkin D, Olasky J, Jones DB, Schwaitzberg SD, Jones SB, Cao CGL, Molina M, Henriques S, Wang J, Flinn J, De S (2017) Or fire virtual training simulator: design and face validity. *Surg Endosc* 31(9):3527–3533. <https://doi.org/10.1007/s00464-016-5379-7>
- Druzhinina MV, Zakharchenko MV, Ananina MA, Solovyova EE (2019) Simulation based technologies for professional training of managers in the arctic region. *Inst Phys Conf Ser Mater Sci Eng* 483(1):012078. <https://doi.org/10.1088/1757-899X/483/1/0120786>
- Dzeng RJ, Hsueh HH, Chang RN (2016) 3d game-based training system for hazard identification on construction site. In: 2015 12th International conference on fuzzy systems and knowledge discovery (fskd 2015) (pp 2453–2458) IEEE. <https://doi.org/10.1109/FSKD.2015.7382339>
- Edwards TC, Patel A, Szyszka B, Coombs AW, Liddle AD, Kucheria R, Cobb JP, Logishetty K (2021) Immersive virtual reality enables technical skill acquisition for scrub nurses in complex revision total knee arthroplasty. *Arch Orthop Trauma Surg* 141(12):2313–2321. <https://doi.org/10.1007/s00402-021-04050-4>
- Eiris R, Jain A, Gheisari M, Wehle A (2020) Safety immersive storytelling using narrated 360-degree panoramas: a fall hazard training within the electrical trade context. *Saf Sci* 127:104703. <https://doi.org/10.1016/j.ssci.2020.104703>
- Erel E, Aiyenibe B, Butler PE (2003) Microsurgery simulators in virtual reality: review. *Microsurgery* 23(2):147–152. <https://doi.org/10.1002/micr.10106>
- Jaskiewicz F, Kowalewski D, Starosta K, Cierniak M, Timler D (2020) Chest compressions quality during sudden cardiac arrest scenario performed in virtual reality: a crossover study in a training environment. *Medicine* 99(48):e23374. <https://doi.org/10.1097/MD.00000000000023374>
- Feng Z, González VA, Amor R, Lovreglio R, Cabrera-Guerrero G (2018) Immersive virtual reality serious games for evacuation training and research: a systematic literature review. *Comput Educ* 127:252–266. <https://doi.org/10.1016/j.compedu.2018.09.002>
- Francone A, Huang JM, Ma J, Tsao TC, Rosen J, Hubschman JP (2019) The effect of haptic feedback on efficiency and safety during preretinal membrane peeling simulation. *Trans vis Sci Technol* 8(4):2. <https://doi.org/10.1167/tvst.8.4.2>
- Freina L, Ott M (2015) A literature review on immersive virtual reality in education: State of the art and perspectives. In: Proceedings of the 11th international scientific conference elearning and software for education (Vol 1, pp 133–141) Carol I National Defence University Publishing House. <https://doi.org/10.12753/2066-026X-15-020>
- Frendo M, Frithioff A, Konge L, Sorensen MS, Andersen SAW (2021) Cochlear implant surgery: learning curve in virtual reality simulation training and transfer of skills to a 3d-printed temporal bone - a prospective trial. *Cochlear Implants Int* 22(6):330–337. <https://doi.org/10.1080/14670100.2021.1940629>
- Fu SX, Liu X, Zhou L, Zhou MS, Wang LM (2017) Applied research on laparoscopic simulator in the resident surgical laparoscopic operation technical training. *Indian J Surg* 79(4):288–293. <https://doi.org/10.1007/s12262-016-1468-z>
- Fu Y, Li Q (2020) Spatial reference frame based user interface design in the virtual reality game design. In: Dmsviva 2020—proceedings of the 26th international dms conference on visualization and visual languages (pp 34–36) KSI Research. <https://doi.org/10.18293/DMSVIVA20-020>
- Ganier F, Hoareau C, Tisseau J (2014) Evaluation of procedural learning transfer from a virtual environment to a real situation: A case study on tank maintenance training. *Ergonomics* 57(6):828–843. <https://doi.org/10.1080/00140139.2014.899628>
- Gao Y, Gonzalez VA, Yiu TW (2019) The effectiveness of traditional tools and computer-aided technologies for health and safety training in the construction sector: a systematic review. *Comput Educ* 138:101–115. <https://doi.org/10.1016/j.compedu.2019.05.003>
- García AA, Bobadilla IG, Figueroa GA, Ramírez MP, Román JM (2016) Virtual reality training system for maintenance and operation of high-voltage overhead power lines. *Virtual Real* 20(1):27–40. <https://doi.org/10.1007/s10055-015-0280-6>
- Gawecki W, Wegrzyniak M, Mickiewicz P, Gawłowska MB, Talar M, Wierzbicka M (2020) The impact of virtual reality training on the quality of real antromastoidectomy performance. *J Clin Med* 9(10):3197. <https://doi.org/10.3390/jcm9103197>
- Gibson ME, Kim JJJ, McManus M, Harris LR (2020) The effect of training on the perceived approach angle in visual vertical heading judgements in a virtual environment. *Exp Brain Res* 238(9):1861–1869. <https://doi.org/10.1007/s00221-020-05841-8>
- Gupta A, Varghese K (2020) Scenario-based construction safety training platform using virtual reality. In: Proceedings of the 37th international symposium on automation and robotics in construction (isarcon 2020) (pp 892–899) The International Association for Automation and Robotics in Construction. <https://doi.org/10.22260/ISARC2020/0123>
- Gurusamy K, Aggarwal R, Palanivelu L, Davidson BR (2008) Systematic review of randomized controlled trials on the effectiveness

- of virtual reality training for laparoscopic surgery. *Br J Surg* 95(9):1088–1097. <https://doi.org/10.1002/bjs.6344>
- Hauschild J, Rivera JC, Johnson AE, Burns TC, Roach CJ (2021) Shoulder arthroscopy simulator training improves surgical procedure performance: a controlled laboratory study. *Orthopaed J Sports Med.* 9(5):23259671211003870. <https://doi.org/10.1177/23259671211003873>
- Havemann MC, Dalsgaard T, Sørensen JL, Røssaak K, Brisling S, Mosgaard BJ, Høgdall C, Bjerrum F (2019) Examining validity evidence for a simulation-based assessment tool for basic robotic surgical skills. *J Robot Surg* 13(1):99–106. <https://doi.org/10.1007/s11701-018-0811-8>
- Herrington A, Tacy J (2020) Crossing the power line: Using virtual simulation to prepare the first responders of utility linemen. *Informatics* 7(3):26. <https://doi.org/10.3390/INFORMATIC57030026>
- Hill RW Jr, Gratch J, Marsella S, Rickel J, Swartout WR, Traum DR (2003) Virtual humans in the mission rehearsal exercise system. *Künstliche Intell* 17(4):5
- Holton EF III (1996) The flawed four-level evaluation model. *Hum Resour Dev Q* 7(1):5–21. <https://doi.org/10.1002/hrdq.3920070103>
- Hoogenes J, Wong N, Al-Harbi B, Kim KS, Vij S, Bolognone E, Quantz M, Guo Y, Shayegan B, Matsumoto ED (2018) A randomized comparison of 2 robotic virtual reality simulators and evaluation of trainees' skills transfer to a simulated robotic urethrovesical anastomosis task. *Urology* 111:110–115. <https://doi.org/10.1016/j.urology.2017.09.023>
- Hsu EB, Li Y, Bayram JD, Levinson D, Yang S, Monahan C (2013) State of virtual reality based disaster preparedness and response training. *PLoS Currents*. <https://doi.org/10.1371/currents.dis.1ea2b2e71237d5337fa53982a38b2aff>
- International Labour Organization (n.d.a) Industries and sectors. <https://www.ilo.org/global/industries-and-sectors/lang--en/index.htm>. Accessed 11 Jan 2023
- Jeelani I, Han K, Albert A (2020) Development of virtual reality and stereo-panoramic environments for construction safety training. *Eng Constr Archit Manag* 27(8):1853–1876. <https://doi.org/10.1108/ECAM-07-2019-0391>
- Jeelani I, Han K, Albert A (2017) Development of immersive personalized training environment for construction workers. In Lin KY, ElGohary N, Tang P (ed) *Computing in civil engineering 2017: Sensing, simulation, and visualization* (pp 407–415) American Society of Civil Engineers. <https://doi.org/10.1061/9780784480830.050>
- Jensen L, Konradsen F (2018) A review of the use of virtual reality head-mounted displays in education and training. *Educ Inf Technol* 23(4):1515–1529. <https://doi.org/10.1007/s10639-017-9676-0>
- Jensen UJ, Jensen J, Ahlberg G, Tornvall P (2016) Virtual reality training in coronary angiography and its transfer effect to real-life catheterisation lab. *EuroIntervention* 11(13):1503–1510. https://doi.org/10.4244/EIJY15M06_05
- Joda T, Gallucci GO, Wismeijer D, Zitzmann NU (2019) Augmented and virtual reality in dental medicine: A systematic review. *Comput Biol Med* 108:93–100. <https://doi.org/10.1016/j.combiomed.2019.03.012>
- Joshi S, Hamilton M, Warren R, Faucett D, Tian W, Wang Y, Ma J (2021) Implementing virtual reality technology for safety training in the precast/ prestressed concrete industry. *Appl Ergon* 90:103286. <https://doi.org/10.1016/j.apergo.2020.103286>
- Jung JK, Ahn YJ (2018) Effects of interface on procedural skill transfer in virtual training: Lifeboat launching operation study. *Comput Animat Virt Worlds* 29(3–4):e1812. <https://doi.org/10.1002/cav.1812>
- Jung JK, Kim HT (2017) Designing effective virtual training: a case study in maritime safety. *J Ergon Soc Korea* 36(5):385–394. <https://doi.org/10.5143/jesk.2017.36.5.385>
- Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *Int J Aviat Psychol* 3(3):203–220
- Keßler D, Arntz A, Friedhoff J, Eimler SC (2020) Mill instructor: Teaching industrial cnc procedures using virtual reality. In 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (pp 231–234) <https://doi.org/10.1109/aiivr50618.2020.00048>
- Kim J, Ahn S, Han S (2020) Impact of vr settings on immersiveness in vr-based construction skills training: Case study of scaffolding work. In: Tang P, Grau D, El Asmar M (ed) *Construction research congress 2020: Computer applications* (pp 336–344) American Society of Civil Engineers. <https://doi.org/10.1061/9780784482865.036>
- Kirkpatrick DL, Kirkpatrick JD (2007) *Implementing the four levels: a practical guide for effective evaluation of training programs* (1st ed) Berrett-Koehler Publishers
- Kirkpatrick DL (1976) Evaluation of training. In Craig RL (ed) *Training and development handbook*. McGraw-Hill
- Knowles MS (1970) *The modern practice of adult education: androgogy versus pedagogy*. New York Association Press
- Kraiger K (2002) Decision-based evaluation. In: Kraiger K (ed) *Creating, implementing, and managing effective training and development*. Jossey-Bass, pp 331–375
- Kraiger K, Ford JK, Salas E (1993) Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *J Appl Psychol* 78(2):311–328. <https://doi.org/10.1037/0021-9010.78.2.311>
- Kwegyir-Afful E, Kantola J (2021) Simulation-based safety training for plant maintenance in virtual reality. In *Advances in intelligent systems and computing* (Vol 1206 pp 167–173) Springer Nature. https://doi.org/10.1007/978-3-030-51064-0_22
- Kwon C (2020) The effect of the degree of anxiety of learners during the use of vr on the flow and learning effect. *Appl Sci* 10(14):4932. <https://doi.org/10.3390/app10144932>
- Lacko J (2020) Health safety training for industry in virtual reality. In *Proceedings of the 30th international conference on cybernetics and informatics, k and i 2020*. <https://doi.org/10.1109/KI48306.2020.9039854>
- Lang Y, Wei L, Xu F, Zhao Y, Yu L (2018) Synthesizing personalized training programs for improving driving habits via virtual reality. In 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp 297–304) IEEE. <https://doi.org/10.1109/VR.2018.8448290>
- Leder J, Horlitz T, Puschmann P, Wittstock V, Schütz A (2019) Comparing immersive virtual reality and powerpoint as methods for delivering safety training: Impacts on risk perception, learning, and decision making. *Saf Sci* 111:271–286. <https://doi.org/10.1016/j.ssci.2018.07.021>
- Li H, Lu M, Chan G, Skitmore M (2015) Proactive training system for safe and efficient precast installation. *Autom Constr* 49:163–174. <https://doi.org/10.1016/j.autcon.2014.10.010>
- Li X, Yi W, Chi H-L, Wang X, Chan AP (2018) A critical review of virtual and augmented reality (vr/ar) applications in construction safety. *Autom Constr* 86:150–162. <https://doi.org/10.1016/j.autcon.2017.11.003>
- Li C, Liang W, Quigley C, Zhao Y, Yu LF (2017) Earthquake safety training through virtual drills. *IEEE Trans vis Comput Graph* 23(4):1275–1284. <https://doi.org/10.1109/TVCG.2017.2656958>
- Liang H, Liang F, Wu F, Wang C, Chang J (2018) Development of a vr prototype for enhancing earthquake evacuee safety. In: *Proceedings—vrcai 2018: 16th acm siggraph international conference on virtual-reality continuum and its applications in industry*.

- Association for Computing Machinery. <https://doi.org/10.1145/3284398.3284417>
- Liang H, Ge C, Liang F, Sun Y, Li P, Wang C (2019a) Training model of safe escape from fire based on virtual reality. In: Proceedings of the 2019a 8th international conference on networks, communication and computing (pp 168–175) Association for Computing Machinery. <https://doi.org/10.1145/3375998.3376013>
- Liang Z, Zhou K, Gao K (2019b) Development of virtual reality serious game for underground rock-related hazards safety training. IEEE Access 7:118639–118649 Article 8795446. <https://doi.org/10.1109/ACCESS.2019.2934990>
- Liang H, Ge C, Liang F, Sun Y (2020) Vr-based training model for enhancing fire evacuee safety. Int J Perform Eng 1:107–117. <https://doi.org/10.23940/ijpe.20.01.p12.107117>
- Liu X, Liu Y, Zhu X, An M, Hu F (2016) Virtual reality based navigation training for astronaut moving in a simulated space station. In: Lecture notes in computer science (Vol 9740 pp 416–423) Springer Nature. https://doi.org/10.1007/978-3-319-39907-2_40
- Logishetty K, Rudran B, Cobb JP (2019) Virtual reality training improves trainee performance in total hip arthroplasty: a randomized controlled trial. Bone Joint J 101B(12):1585–1592. <https://doi.org/10.1302/0301-620x.101b12.Bjj-2019-0643.R1>
- LoJacono CT, MacPherson RP, Kuznetsov NA, Raisbeck D, Ross SE, Rhea CK (2018) Obstacle crossing in a virtual environment transfers to a real environment. J Motor Learn Dev 6(2):234–249. <https://doi.org/10.1123/jml.2017-0019>
- Lovreglio R, Duan X, Rahouti A, Phipps R, Nilsson D (2021) Comparing the effectiveness of fire extinguisher virtual reality and video training. Virtual Reality 25(1):133–145. <https://doi.org/10.1007/s10055-020-00447-5>
- Makransky G, Borre-Gude S, Mayer RE (2019) Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. J Comput Assist Learn 35(6):691–707. <https://doi.org/10.1111/jcal.12375>
- McBain-Rigg KE, Franklin RC, King JC, Lower T (2017) Influencing safety in Australian agriculture and fisheries. J Agromed 22(4):347–357. <https://doi.org/10.1080/1059924X.2017.1353935>
- Merchant Z, Goetz ET, Cifuentes L, Keeney-Kennicutt W, Davis TJ (2014) Effectiveness of virtual reality-based instruction on students' learning outcomes in k-12 and higher education: a meta-analysis. Comput Educ 70:29–40. <https://doi.org/10.1016/j.compedu.2013.07.033>
- Mirauda D, Capece N, Erra U (2020) Sustainable water management: Virtual reality training for open-channel flow monitoring. Sustain Switz 12(3):757. <https://doi.org/10.3390/su12030757>
- Mondragón-Bernal IF (2020) A system for immersive medical and engineering training based on serious games. Ing Y Univ 24:1–33. <https://doi.org/10.11144/javeriana.ined24.sime>
- Moore HF, Eiris R, Gheisari M, Esmaili B (2019) Hazard identification training using 360-degree panorama vs. Virtual reality techniques: A pilot study. In: Cho YK, Leite F, Behzadan A, Wang C (ed) Computing in civil engineering 2019: Visualization, information modeling, and simulation (pp 55–62) American Society of Civil Engineers. <https://doi.org/10.1061/9780784482421.008>
- Nagendran M, Gurusamy KS, Aggarwal R, Loizidou M, Davidson BR (2013) Virtual reality training for surgical trainees in laparoscopic surgery. Cochrane Database Syst Rev. <https://doi.org/10.1002/14651858.CD006575.pub3>
- Narciso D, Melo M, Rodrigues S, Paulo Cunha J, Vasconcelos-Raposo J, Bessa M (2021) A systematic review on the use of immersive virtual reality to train professionals. Multimed Tools Appl 80(9):13195–13214. <https://doi.org/10.1007/s11042-020-10454-y>
- National Research Council (1995) Virtual reality: scientific and technological challenges. Nat Acad Press. <https://doi.org/10.17226/4761>
- Neis F, Brucker S, Henes M, Taran FA, Hoffmann S, Wallwiener M, Schönfisch B, Ziegler N, Larbig A, De Wilde RL (2016) Evaluation of the hystsim™-virtual reality trainer: an essential additional tool to train hysteroscopic skills outside the operation theater. Surg Endosc 30(11):4954–4961. <https://doi.org/10.1007/s00464-016-4837-6>
- Next World Enterprises (2022) VR Training Courses. <https://nextworldenterprises.com/vr-training/>
- Nilsson C, Sorensen JL, Konge L, Westen M, Stadeager M, Ottesen B, Bjerrum F (2017) Simulation-based camera navigation training in laparoscopy—a randomized trial. Surg Endosc 31(5):2131–2139. <https://doi.org/10.1007/s00464-016-5210-5>
- Nykänen M, Puro V, Tiikkaja M, Kannisto H, Lantto E, Simpura F, Uusitalo J, Lukander K, Räsänen T, Heikkilä T, Teperi AM (2020) Implementing and evaluating novel safety training methods for construction sector workers: Results of a randomized controlled trial. J Saf Res 75:205–221. <https://doi.org/10.1016/j.jsr.2020.09.015>
- Oh SM, Kim JY, Han S, Lee W, Kim I, Hong G, Oh W, Moon H, Seo C (2020) Development and usability of a virtual reality-based filler injection training system. Aesthetic Plast Surg 44(5):1833–1842. <https://doi.org/10.1007/s00266-020-01872-2>
- Orland MD, Patetta MJ, Wieser M, Kayupov E, Gonzalez MH (2020) Does virtual reality improve procedural completion and accuracy in an intramedullary tibial nail procedure? A randomized control trial. Clin Orthop Relat Res 478(9):2170–2177. <https://doi.org/10.1097/corr.0000000000001362>
- Ouyang SG, Wang G, Yao JY, Zhu GHW, Liu ZY, Feng C (2018) A unity3d-based interactive three-dimensional virtual practice platform for chemical engineering. Comput Appl Eng Educ 26(1):91–100. <https://doi.org/10.1002/cae.21863>
- Page RL (2000) Brief history of flight simulation. In: SimTecT 2000 Proceedings pp 11–17
- Pedram S, Palmisano S, Skarbez R, Perez P, Farrelly M (2020) Investigating the process of mine rescuers' safety training with immersive virtual reality: a structural equation modelling approach. Comput Educ 153:103891. <https://doi.org/10.1016/j.compedu.2020.103891>
- Pelargos PE, Nagasawa DT, Lagman C, Tenn S, Demos JV, Lee SJ, Bui TT, Barnette NE, Bhatt NS, Ung N, Bari A, Martin NA, Yang I (2017) Utilizing virtual and augmented reality for educational and clinical enhancements in neurosurgery. J Clin Neurosci 35:1–4. <https://doi.org/10.1016/j.jocn.2016.09.002>
- Pellas N, Mystakidis S, Kazanidis I (2021) Immersive virtual reality in k-12 and higher education: a systematic review of the last decade scientific literature. Virtual Real 25(3):835–861. <https://doi.org/10.1007/s10055-020-00489-9>
- Pham HC, Dao NN, Kim JU, Cho S, Park CS (2018) Energy-efficient learning system using web-based panoramic virtual photorealism for interactive construction safety education. Sustainability 10(7):2262. <https://doi.org/10.3390/su10072262>
- Phillips P (2003) Training evaluation in the public sector. Dissertation, The University of Southern Mississippi
- Piromchai P, Avery A, Laopaiboon M, Kennedy G, O'Leary S (2015) Virtual reality training for improving the skills needed for performing surgery of the ear, nose or throat. Cochrane Database Syst Rev. <https://doi.org/10.1002/14651858.CD010198.pub2>
- Pixo VR (n.d.a) Virtual Reality for Construction Training. <https://pixovr.com/vr-training-categories/construction/>. Accessed 11 Jan 2023
- Polivka BJ, Anderson S, Darragh AR, Lavender SA, Sommerich CM, Stredney DL, Wills CE (2019) Efficacy and usability of a virtual simulation training system for health and safety hazards encountered by healthcare workers. Games Health J 8(2):121–128. <https://doi.org/10.1089/g4h.2018.0068>

- Prasolova-Førland E, Molka-Danielsen J, Fominykh M, Lamb K (2017) Active learning modules for multi-professional emergency management training in virtual reality. In: 2017 IEEE 6th international conference on teaching, assessment, and learning for engineering (tale) (pp 461–468) IEEE. <https://doi.org/10.1109/TALE.2017.8252380>
- Prendinger H, Alvarez N, Sanchez-Ruiz A, Cavazza M, Catarino J, Oliveira J, Prada R, Fujimoto S, Shigematsu M (2016) Intelligent biohazard training based on real-time task recognition. *ACM Trans Interact Intell Syst* 6(3):21. <https://doi.org/10.1145/2883617>
- Radhakrishnan U, Koumaditis K, Chinello F (2021) A systematic review of immersive virtual reality for industrial skills training. *Behav Inform Technol* 40(12):1310–1339. <https://doi.org/10.1080/0144929X.2021.1954693>
- Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I (2020) A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. *Comput Educ* 147:103778. <https://doi.org/10.1016/j.compedu.2019.103778>
- Rahouti A, Lovreglio R, Datoussaïd S, Descamps T (2021) Prototyping and validating a non-immersive virtual reality serious game for healthcare fire safety training. *Fire Technol* 57(6):3041–3078. <https://doi.org/10.1007/s10694-021-01098-x>
- Reio TG, Rocco TS, Smith DH, Chang E (2017) A critique of Kirkpatrick's evaluation model. *New Horizons Adult Educ Human Resour Dev* 29(2):35–53. <https://doi.org/10.1002/nha3.20178>
- Renganayagalu SK, Mallam SC, Nazir S (2021) Effectiveness of vr head mounted displays in professional training: a systematic review. *Technol Knowl Learn* 26:999–1041. <https://doi.org/10.1007/s10758-020-09489-9>
- Ricci F, Chiesi A, Bisio C, Panari C, Pelosi A (2016) Effectiveness of occupational health and safety training: a systematic review with meta-analysis. *J Work Learn* 28(6):355–377. <https://doi.org/10.1108/JWL-11-2015-0087>
- Robson LS, Stephenson CM, Schulte PA, Amick BC, Irvin EL, Eggerth DE, Chan S, Bielecky AR, Wang AM, Heidotting TL, Peters RH, Clarke JA, Cullen K, Rotunda CJ, Grubb PL (2012) A systematic review of the effectiveness of occupational health and safety training. *Scand J Work Environ Health* 38(3):193–208. <https://doi.org/10.5271/sjweh.3259>
- Rose FD, Attree EA, Brooks BM, Parslow DM, Penn PR (2000) Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics* 43(4):494–511. <https://doi.org/10.1080/001401300184378>
- Rosler KL, Sankaranarayanan G, Duvall A (2019) Acquisition of fire safety knowledge and skills with virtual reality simulation. *Nurse Educ* 44(2):88–92. <https://doi.org/10.1097/NNE.0000000000000551>
- Sacks R, Perlman A, Barak R (2013) Construction safety training using immersive virtual reality. *Constr Manag Econ* 31(9):1005–1017. <https://doi.org/10.1080/01446193.2013.828844>
- Safe Work Australia (2016a) Work-related traumatic injury fatalities, australia 2015. <https://www.safeworkaustralia.gov.au/system/files/documents/1702/work-related-traumatic-injury-fatalities.pdf>
- Safe Work Australia (2016b) Work health and safety in the agricultural industry 2016b. <https://www.safeworkaustralia.gov.au/system/files/documents/1702/whs-in-the-agricultural-industry.pdf>
- Safe Work Australia (2021) Work-related injury fatalities—key whs statistics australia 2021. <https://www.safeworkaustralia.gov.au/sites/default/files/2021-10/Key%20work%20health%20and%20safety%20statistics%20Australia%202021.pdf>
- Saghafian M, Laumann K, Akhtar RS, Skogstad MR (2020) The evaluation of virtual reality fire extinguisher training. *Front Psychol* 11:593466. <https://doi.org/10.3389/fpsyg.2020.593466>
- Salas E, Tannenbaum SI, Kraiger K, Smith-Jentsch KA (2012) The science of training and development in organizations: what matters in practice. *Psychol Sci Public Interest* 13(2):74–101. <https://doi.org/10.1177/1529100612436661>
- Salcedo JN, Lackey SJ, Maraj C (2016) Impact of instructional strategies on workload, stress, and flow in simulation-based training for behavior cue analysis. In: *Lecture notes in computer science* (Vol 9740 pp 184–195) Springer Nature. https://doi.org/10.1007/978-3-319-39907-2_18
- Sankaranarayanan G, Li B, Miller A, Wakily H, Jones SB, Schwaitzberg S, Jones DB, De S, Olasky J (2016) Face validation of the virtual electrosurgery skill trainer (vest©). *Surg Endosc* 30(2):730–738. <https://doi.org/10.1007/s00464-015-4267-x>
- Satava RM (1995) Medical applications of virtual reality. *J Med Syst* 19(3):275–280. <https://doi.org/10.1007/BF02257178>
- Schubert T, Friedmann F, Regenbrecht H (2001) The experience of presence: Factor analytic insights. *Presence Teleoperators Virtual Environ* 10(3):266–281. <https://doi.org/10.1162/10547460100343603>
- Schulz GB, Grimm T, Buchner A, Jokisch F, Casuscelli J, Kretschmer A, Mumm JN, Ziegelmeüller B, Stief CG, Karl A (2019) Validation of a high-end virtual reality simulator for training transurethral resection of bladder tumors. *J Surg Educ* 76(2):568–577. <https://doi.org/10.1016/j.jsurg.2018.08.001>
- Sebastian K, Ghose T, Huff M (2018) Repeating virtual assembly training facilitates memory for coarse but not fine assembly steps. *J Comput Assist Learn* 34(6):787–798. <https://doi.org/10.1111/jcal.12285>
- Shannon HS, Robson LS, Guastello SJ (1999) Methodological criteria for evaluating occupational safety intervention research. *Saf Sci* 31(2):161–179. [https://doi.org/10.1016/S0925-7535\(98\)00063-0](https://doi.org/10.1016/S0925-7535(98)00063-0)
- Sharma S, Bodempudi ST, Scribner D, Grazaitis P (2020) Active shooter response training environment for a building evacuation in a collaborative virtual environment. *Int Symp Electron Imag Sci Technol* 13:223. <https://doi.org/10.2352/ISSN.2470-1173.2020.13.ERVR-223>
- Shewaga R, Uribe-Quevedo A, Kapralos B, Lee K, Alam F (2018) A serious game for anesthesia-based crisis resource management training. *Comput Entertain* 16(2):6. <https://doi.org/10.1145/3180660>
- Shi Y, Du J, Ahn CR, Ragan E (2019) Impact assessment of reinforced learning methods on construction workers' fall risk behavior using virtual reality. *Autom Constr* 104:197–214. <https://doi.org/10.1016/j.autcon.2019.04.015>
- Shi Y, Du J, Zhu Q, Liu X (2020) The impact of engineering information formats on workers' cognitive load in working memory development. In: El Asmar M, Tang P, Grau D (ed) *Construction research congress 2020: Infrastructure systems and sustainability* (pp 79–88) American Society of Civil Engineers. <https://doi.org/10.1061/9780784482858.010>
- Shim JS, Noh TI, Kim JY, Pyun JH, Cho S, Oh MM, Kang SH, Cheon J, Lee JG, Kim JJ, Kang SG (2018) Predictive validation of a robotic virtual reality simulator: the tube 3 module for practicing vesicourethral anastomosis in robot-assisted radical prostatectomy. *Urology* 122:32–36. <https://doi.org/10.1016/j.urology.2018.08.013>
- Sim ZH, Chook Y, Hakim MA, Lim WN, Yap KM (2019) Design of virtual reality simulation-based safety training workshop. In 2019 IEEE international symposium on haptic, audio and visual environments and games (have) (pp 1–6) IEEE. <https://doi.org/10.1109/HAVE.2019.8921366>
- Simpson TG, Rafferty K (2020) Evaluating the effect of reinforcement haptics on motor learning and cognitive workload in

- driver training. In: Lecture notes in computer science (Vol 12242 pp 203–211) Springer Nature. https://doi.org/10.1007/978-3-030-58465-8_16
- Sinitsky DM, Fernando B, Potts H, Lykoudis P, Hamilton G, Berlingieri P (2020) Development of a structured virtual reality curriculum for laparoscopic appendicectomy. *Am J Surg* 219(4):613–621. <https://doi.org/10.1016/j.amjsurg.2019.04.020>
- Slater M (2009) Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos Trans R Soc B Biol Sci* 364(1535):3549–3557. <https://doi.org/10.1098/rstb.2009.0138>
- Slater M, Sanchez-Vives MV (2016) Enhancing our lives with immersive virtual reality. *Front Robot AI*. <https://doi.org/10.3389/frobt.2016.00074>
- Smith J, Doody K, Veitch B (2019) Being prepared for emergencies: a virtual environment experiment on the retention and maintenance of egress skills. *WMU J Marit Aff* 18(3):425–449. <https://doi.org/10.1007/s13437-019-00174-y>
- Smith J, Veitch B (2019) A better way to train personnel to be safe in emergencies. *J Risk Uncertain Eng Syst Part B Mech Eng* 5(1):011003. <https://doi.org/10.1115/1.4040660>
- Song H, Kim T, Kim J, Ahn D, Kang Y (2021) Effectiveness of vr crane training with head-mounted display: double mediation of presence and perceived usefulness. *Autom Construct* 122:103506. <https://doi.org/10.1016/j.autcon.2020.103506>
- Sportillo D, Paljic A, Ojeda L (2018) Get ready for automated driving using virtual reality. *Accid Anal Prev* 118:102–113. <https://doi.org/10.1016/j.aap.2018.06.003>
- Sugand K, Mawkin M, Gupte C (2016) Training effect of using touch surgery™ for intramedullary femoral nailing. *Injury* 47(2):448–452. <https://doi.org/10.1016/j.injury.2015.09.036>
- Suh A, Prophet J (2018) The state of immersive technology research: a literature analysis. *Comput Hum Behav* 86:77–90. <https://doi.org/10.1016/j.chb.2018.04.019>
- Suto H, Zhang X, Shen X, Raksinchareonsak P, Tsumura N (2020) Development and evaluation of immersive educational system to improve driver's risk prediction ability in traffic accident situation. *Is & T Int Sympos Electron Imag Sci Technol* 13:380. <https://doi.org/10.2352/ISSN.2470-1173.2020.13.ERVR-380>
- Ta F, Valentine A, Male S, Hassan GM (2019) Piloted online training module to teach on-site safety in engineering. In: Chew YW, Chan KM, Alphonso A (ed) Personalised learning. Diverse goals. One heart. Ascilite 2019 singapore (pp 292–301)
- Tang YM, Ng GWY, Chia NH, So EHK, Wu CH, Ip WH (2021) Application of virtual reality (vr) technology for medical practitioners in type and screen (t & s) training. *J Comput Assist Learn* 37(2):359–369. <https://doi.org/10.1111/jcal.12494>
- Tannenbaum S, Methieu J, Cannon-Bowers J, Salas E (1993) Factors that influence training effectiveness: A conceptual model and longitudinal analysis (Report No 93–001) Naval Training Systems Center. <https://apps.dtic.mil/sti/pdfs/ADA281519.pdf>
- Tawadrous M, Rojas D, Kapralos B, Hogue A, Dubrowski A (2017) The effects of stereoscopic 3d on knowledge retention within a serious gaming environment. *Multimed Tools Appl* 76(5):7301–7319. <https://doi.org/10.1007/s11042-016-3394-2>
- Tichon J, Burgess-Limerick R (2011) A review of virtual reality as a medium for safety related training in mining. *J Health Saf Res Pract* 3(1):33–40
- Torres-Guerrero F, Neira-Tovar L, Torres-Treviño L (2019) An introductory training for welding workshop: A biometric evaluation using virtual reality scenes to improve practice. In: Lecture notes in mechanical engineering (pp 319–331) Springer Nature. https://doi.org/10.1007/978-3-030-18715-6_27
- Truong H, Qi D, Ryason A, Sullivan AM, Cudmore J, Alfred S, Jones SB, Parra JM, De S, Jones DB (2021) Does your team know how to respond safely to an operating room fire? Outcomes of a virtual reality, ai-enhanced simulation training. *Surg Endosc*. <https://doi.org/10.1007/s00464-021-08602-y>
- Tsuboi H, Toyama S, Nakajima T (2018) Enhancing bicycle safety through immersive experiences using virtual reality technologies. In: Lecture notes in computer science (Vol 10915 pp 444–456) Springer Nature. https://doi.org/10.1007/978-3-319-91470-1_35
- U. S. General Services Administration (n.d.a) Usability evaluation basics. <https://www.usability.gov/what-and-why/usability-evaluation.html#:~:text=What%20is%20Usability%3F,overall%20satisfaction%20of%20the%20user.>
- Usoh M, Catena E, Arman S, Slater M (2000) Using presence questionnaires in reality. *Presence Teleoper Virtual Environ* 9(5):497–503. <https://doi.org/10.1162/105474600566989>
- Vahdatikhaki F, El Ammari K, Langroodi AK, Miller S, Hammad A, Doree A (2019) Beyond data visualization: a context-realistic construction equipment training simulators. *Autom Construct* 106:102853. <https://doi.org/10.1016/j.autcon.2019.102853>
- Vaughan N, Dubey VN, Wainwright TW, Middleton RG (2016) A review of virtual reality based training simulators for orthopaedic surgery. *Med Eng Phys* 38(2):59–71. <https://doi.org/10.1016/j.medengphy.2015.11.021>
- Worksafe Victoria (2022) Industries. <https://www.worksafe.vic.gov.au/industries>
- Wang Y, Guo S, Li Y, Tamiya T, Song Y (2018) Design and evaluation of safety operation vr training system for robotic catheter surgery. *Med Biol Eng Comput* 56(1):25–35. <https://doi.org/10.1007/s11517-017-1666-2>
- Wang KC, Bernardoni ED, Cotter EJ, Levine BR, Frank RM (2021) The impact of unguided trauma simulation practice on novice performance: a randomized controlled trial. *J Am Acad Orthop Surg* 29(6):255–262. <https://doi.org/10.5435/jaaos-d-19-00225>
- Waterman BR, Martin KD, Cameron KL, Owens BD, Belmont PJ (2016) Simulation training improves surgical proficiency and safety during diagnostic shoulder arthroscopy performed by residents. *Orthopedics* 39(3):e479–e485. <https://doi.org/10.3928/01477447-20160427-02>
- Weber A, Nickel P, Hartmann U, Friemert D, Karamanidis K (2020) Contributions of training programs supported by vr techniques to the prevention of stf accidents. In: Lecture notes in computer science (Vol 12198 pp 276–290) Springer Nature. https://doi.org/10.1007/978-3-030-49904-4_20
- Witmer BG, Singer MJ (1998) Measuring presence in virtual environments: a presence questionnaire. *Presence Teleoper Virtual Environ* 7(3):225–240. <https://doi.org/10.1162/105474698565686>
- Wu HT, Yu WD, Gao RJ, Wang KC, Liu KC (2020a) Measuring the effectiveness of vr technique for safety training of hazardous construction site scenarios. In 2nd IEEE International conference on architecture, construction, environment and hydraulics 2020a (icaceh) (pp 36–39) IEEE. <https://doi.org/10.1109/ICACEH51803.2020.9366218>
- Wu SH, Huang CC, Huang SS, Yang YY, Liu CW, Shulruf B, Chen CH (2020b) Effect of virtual reality training to decreases rates of needle stick/sharp injuries in new-coming medical and nursing interns in taiwan. *J Educ Evaluation Health Profess* 17:1. <https://doi.org/10.3352/jeehp.2020.17.1>
- Xiao X, Zhao S, Meng Y, Soghier L, Zhang X, Hahn J (2020) A physics-based virtual reality simulation framework for neonatal endotracheal intubation. In 2020 IEEE Conference on virtual reality and 3d user interfaces (vr) (pp 557–565) IEEE. <https://doi.org/10.1109/vr46266.2020.00-27>
- Xin B, Chen G, Wang Y, Bai G, Gao X, Chu J, Xiao J, Liu T (2019) The efficacy of immersive virtual reality surgical simulator training for pedicle screw placement: a randomized double-blind

- controlled trial. *World Neurosurg* 124:E324–E330. <https://doi.org/10.1016/j.wneu.2018.12.090>
- Xin B, Huang X, Wan W, Lv K, Hu Y, Wang J, Li S, Zou W, Xiao J, Liu T (2020) The efficacy of immersive virtual reality surgical simulator training for pedicle screw placement: a randomized double-blind controlled trial. *Int Orthop* 44(5):927–934. <https://doi.org/10.1007/s00264-020-04488-y>
- Xu J, Tang Z, Zhao H, Zhang J, Beale R (2019b) Hand gesture-based virtual reality training simulator for collaboration rescue of a railway accident. *Interact Comput* 31(3):577–588. <https://doi.org/10.1093/iwc/iwz037>
- Xu Z, Zheng N (2021) Incorporating virtual reality technology in safety training solution for construction site of urban cities. *Sustainability* 13(1):243. <https://doi.org/10.3390/su13010243>
- Xu S, Ni Q, Du Q (2019a) The effectiveness of virtual reality in safety training: Measurement of emotional arousal with electromyography. In: *Proceedings of the 36th international symposium on automation and robotics in construction (isarc 2019a)* (pp 20–25) The International Association for Automation and Robotics in Construction. <https://doi.org/10.22260/isarc2019/0003>
- Yang C, Kalinitschenko U, Helmert JR, Weitz J, Reissfelder C, Mees ST (2018) Transferability of laparoscopic skills using the virtual reality simulator. *Surg Endosc* 32(10):4132–4137. <https://doi.org/10.1007/s00464-018-6156-6>
- Yau YW, Li Z, Chua MT, Kuan WS, Chan GWH (2021) Virtual reality mobile application to improve videoscopic airway training a randomised trial. *Ann Acad Med Singapore* 50(2):141–148
- Yiannakopoulou E, Nikiteas N, Perrea D, Tsigris C (2015) Virtual reality simulators and training in laparoscopic surgery. *Int J Surg* 13:60–64. <https://doi.org/10.1016/j.ijso.2014.11.014>
- Yu M, Yang M, Ku B, Mann JS (2021) Effects of virtual reality simulation regarding high-risk neonatal infection control on nursing students. *Asian Nurs Res* 15(3):189–196. <https://doi.org/10.1016/j.anr.2021.03.002>
- Zhang H (2017) Head-mounted display-based intuitive virtual reality training system for the mining industry. *Int J Min Sci Technol* 27(4):717–722. <https://doi.org/10.1016/j.ijmst.2017.05.005>
- Zhang K, Suo J, Chen J, Liu X, Gao L (2017) Design and implementation of fire safety education system on campus based on virtual reality technology. In: *2017 Federated conference on computer science and information systems (fedcsis)* (pp 1297–1300) IEEE. <https://doi.org/10.15439/2017F376>
- Zhao D, McCoy A, Kleiner B, Feng Y (2016) Integrating safety culture into osh risk mitigation: a pilot study on the electrical safety. *J Civ Eng Manag* 22(6):800–807. <https://doi.org/10.3846/13923730.2014.914099>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.