INSTITUTE FOR DEFENSE ANALYSES

Training Effectiveness Framework for Augmented and Virtual Reality: Developing a Knowledge Base

James Belanich
Franklin L. Moses
Emily A. Fedele

December 2022

Distribution A.
Approved for public release;
distribution is unlimited.

IDA Document D-33349

Log: H 22-000520

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305-3086

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001, project AK-2-4836, “Research, Applications and Opportunities for Augmented and Virtual Reality for Military Education and Training,” for the Air Force Research Laboratory, Airman Systems Directorate, Warfighter Interaction and Readiness Division (AFRL/RHW) in support of the Commander Air Combat Command's (COMACC) migration to Proficiency Based Training, and the Office of the Under Secretary of Defense, Research and Engineering. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

James Belanich, Project Leader
jbelanic@ida.org, 703-845-6606

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Copyright Notice

© 2023 Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305-3086 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (Feb. 2014).

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-33349

Training Effectiveness Framework for Augmented and Virtual Reality: Developing a Knowledge Base

James Belanich
Franklin L. Moses
Emily A. Fedele

Executive Summary

The broad assertion of augmented reality (AR) and virtual reality (VR) effectiveness for training and for performance enhancement is overly simplistic and clouds important factors that influence the use and generalizability of findings. Effectiveness depends on factors such as type of AR or VR technology used, the virtual content provided, the user, and the kind of tasks to be done. In addition, studies from disparate disciplines make the generalization of findings more complex by emphasizing different factors. This report is about the development of a framework for organizing studies, its application to a searchable knowledge base, and the kinds of results that may satisfy user needs now and in the future.

Background

AR/VR Training and Performance Enhancement

Prior analytic reviews of AR/VR primarily depended on meta-analyses and narrative reviews. Narrative reviews focus on particular topics of interest to their authors as a subjective way to organize research. Meta-analyses are a systematic way to combine data from multiple studies and can identify common effects or the reasons for variation. Neither fully describes dimensions that influence the use and generalizability of AR/VR effectiveness.

Factors to Consider in AR/VR Effectiveness

While AR/VR can have value for training, it is important to consider what factors and circumstances influence success. It’s generally true that AR/VR compared to alternatives take less time, increase amount learned, and may result in fewer errors. In addition, users report a general preference for AR/VR. However, more focused studies allow us to identify factors that are important to training effectiveness, and how education level of the users and domain of study are moderating factors influencing effectiveness. In short, there are subtleties to determining AR/VR technology’s relevance and successfully incorporating it into training.

Many studies emphasize the technology features, particularly display technologies that can moderate AR/VR effectiveness. In addition, there are other technology features that may benefit training. For example, many AR/VR systems track the user. This includes where the user is looking (e.g., head orientation, eye tracking), what they are doing with their hands, or where they are positioned in an environment. Tracking information like this may be useful for both developing training by understanding how an expert might behave in a situation and how to teach it to a novice. As these few examples illustrate, AR/VR utility is a function of multiple factors.

Part of organizing the field is to develop a framework—a classification system in which components are organized into groups or types based on particular characteristics. The development of this framework supports a systematic method and format allowing designers and other users to access and understand information they need.

AR/VR Framework Methods and Procedures

Components of the Framework

The effectiveness framework features four dimensions and an outcome category represented by specific questions to be answered for each AR/VR study as entries for a knowledge base. The dimensions and the outcome category with associated questions are:

1. Technology – AR/VR system and components being assessed: What is the device or system being used?
2. Skill/Task or domain – features of the task/job the user learns: What is the task being trained or performed?
3. Training-Technology Integration – features of how the technology is used to train: How is the technology used or integrated or trained in a situation/course? What is the training method?
4. Users – description of the trainee's characteristics, the users of the technology: Who are the performers/trainees? What is the prior skill level of the user for the task and with the technology?
5. Outcome – method of assessment: What metrics were used to measure effectiveness? How effective was the AR/VR in supporting performance?

Using the Framework to Develop a Knowledge Base

AR/VR empirical studies were assessed based on the framework and used to populate a Microsoft Access knowledge base.[1] The knowledge base can be queried by targeting specific descriptions (e.g., Microsoft HoloLens; landing an F-14 on a carrier) or more general characteristics (e.g., see-through head-mounted display (HMD); eye-hand coordination). The ultimate goal of the knowledge base is that it can help users probe the evidence of AR/VR effectiveness in a more nuanced and systematic way.

In addition to the narrative descriptions in the framework, each framework category about a study is tagged with specific codes to classify it. For example, technology used in a study could be a visual display on a computer monitor (tag = visual_computer), or an occluded HMD device (tag

[1] Microsoft Access is a relational database management system (RDBMS) that is part of the Microsoft Office suite included in the Professional and higher editions or sold separately.

= visual_HMD_occluded). The tags help label common characteristics of studies in the knowledge base.

Results from the Knowledge Base

The overall result of this effort is an expandable knowledge base, a query-enabled repository of training effectiveness studies. It supports multiple search strategies to find empirical evidence most relevant to a user's specific needs. For example, users may ask a general question like, "Are AR/VR systems effective for training?" that looks across all of the studies. The result would identify studies categorized as more effective (22), mixed (15), or less effective (4) than traditional training. In addition, 23 studies assessed AR/VR system components to determine what mix of features influence effectiveness. This approach allows users to focus on which studies are more or less effective and how changing and refining system components may influence effectiveness.

AR/VR Technology

One approach to the knowledge base is to search for technology components to find technologies with substantial evidence (i.e., used in many studies) and what technologies have little evidence. The most frequent technology in the knowledge base is an occluded HMD. Other visual-display technologies include computer monitors, see-through HMDs, projections on a surface, and handheld tablets. The knowledge base also has non-display technologies such as those that track a user's hands or head, or a user's position. By categorizing studies by technology, the knowledge base can help a user find studies for specific technologies of interest. In addition, knowledge base queries can help identify where gaps in the evidence exist (i.e., few to no studies).

Skill/Task Domains

Within the knowledge base, there are studies that represent many different skill/task training domains. Discussions with potential users indicated that their first knowledge base query might be for skills/tasks like ones they were planning to train. The most frequently trained skills/tasks are maintenance and assembly tasks as well as medical skills, where there have been several demonstrations of how AR/VR technology can be effective. There are also studies on flight training and military tasks, but those domains have fewer studies and the mixed results suggest that this area is still developing how best to use AR/VR technology. By doing searches with generic skill/task descriptors, users can find studies of direct interest or that share some characteristics.

Training-Technology Integration

How an AR/VR system is incorporated into a training situation is another way that the knowledge base can differentiate between studies. The most frequent set of studies was using the AR/VR technology as a practice environment, where users can practice the training task over and over. The next most frequent training use was the overlaying of extra symbols or information on the visual display to provide cues or information to the user. For feedback delivery, two broad types included automated feedback about the user’s performance level and instructor feedback. Additionally, some systems go beyond training in that they are expected to be used on the job as an aid.

User Characteristics

Understanding the initial state of the user (i.e., trainee) prior to the training is important because experience with virtual environments and related technology interacts with subsequent performance. The level of user experience for the task to be trained is an important consideration in understanding if a study is well-structured and relevant. In addition, a description of the participants is needed to decide whether study results may or may not generalize. A considerable gap in the studies is the large number, just a little less than half, that did not provide information on the level of experience of the user.

Assessment Metrics

A knowledge base user also may approach outcomes and their metrics from different perspectives. Six clusters of metric types that appear frequently are: physical performance, ratings of performance (e.g., Likert scale), cognitive performance (e.g., multiple-choice knowledge test), usability (e.g., System Usability Scale), workload (e.g., National Aeronautics and Space Administration (NASA) task load index), and opinion. In the knowledge base, studies often use more than one type of metric to provide a better perspective of a system’s effectiveness. For example, physical performance shows how well a skill was learned while usability and workload effect a user’s motivation to employ a system. In addition, opinions provide recommendations for improvements to the technology particularly when the results of system comparisons are mixed.

Intersection of Framework Components

A powerful benefit of knowledge base queries is that they can cut across framework components that are specified by a user. Queries can not only show the range of technologies included in AR/VR studies, but also can drill down deeper into how a particular technology might be used with a specific training method as part of a training system. For example, knowledge base users can determine how occluded head-mounted visual displays (i.e., a particular technology) is integrated into training with automated feedback. They can query the knowledge base by combining its many different dimensions in cross-walks of technology type, skill/task trained or performed, training use, and users.

Usability Test

The practical utility of the framework and resulting knowledge base depend on usability tests to show how well they work and what changes are needed. The first knowledge elicitation, design, development, and usability testing completed in June 2022. The knowledge base as tested included search, sort, and filtering capabilities. The sample of five candidate end users[2] were active-duty military personnel, Department of Defense (DoD) civilians, and contractors with a mix of work-related roles including instructional designers/developers, graphic designers, game designers, research scientists, software engineers, and domain subject-matter experts.

The users were tasked to: 1) search for and download an article from the knowledge base, and 2) upload and annotate an entry into the knowledge base. As users performed the tasks, they were instructed to talk out loud. The primary criterion measure was the Post-Study System Usability Questionnaire (PSSUQ). The PSSUQ is a 16-item usability scale that measures perceived satisfaction with a system. Questions focus on the quality of the system/process, the information provided by the system, and the interface.

An analysis of results reveals some consistent patterns. First, with the exception of a single item ("The system gave me error messages that clearly told me how to fix the problems"), all of the item means were better than the scale midpoint value of 4.0. Second, the users reported favorable attitudes towards the portal. An analysis of the users' qualitative comments included the desire for greater clarity of terminology that led to changes in the knowledge base.

Conclusion

The current state of AR/VR technology effectiveness studies is fragmented and disorganized. A framework documented in this report helped to organize published information and populate a knowledge base with a set of information categories. This knowledge base enables users to filter and find AR/VR studies and information to address their specific needs. Feedback from usability tests guided improvements to interface design and content in the knowledge base. The result is an initial knowledge base of 64 studies that can grow to provide an ever-improving resource for users to understand how AR/VR can be effective.

[2] Previous research suggests that a sample size of five is sufficient to identify 85% of usability problems (Faulkner 2003).

Contents

1. Introduction

Numerous meta-analyses and narrative reviews suggest that augmented reality (AR) and virtual reality (VR) technologies can be effective for training and for performance enhancement (Fletcher et al. 2017; Kaplan et al. 2021; Garzón & Acevedo 2019; Batdi and Talan 2019). However, the broad assertion of AR and VR effectiveness for training and for performance enhancement is overly simplistic and clouds important factors that influence the use and generalizability of findings. AR and VR use many different technologies and the functional differences across those technologies may vary across multiple factors (e.g., immersion, audio/visual fidelity, field of view, refresh rate). In addition, many different kinds of disciplines and publications report on effectiveness studies. We developed a framework to organize a knowledge base of studies for users to better translate empirical findings into what they need. Chapter Two provides more information about the framework.

The scope and variety of study findings illuminate why compiling them for different uses is difficult. Studies are done in many disparate fields including computer science and engineering that focus on the technology, education that focuses on instructional methods, and specific domains (e.g., medical, military, and equipment maintenance) that focus on implementation in their disciplines. How to distill findings into coherent themes and develop principles across studies posed multifaceted challenges. By using a framework's standardized structure to make AR/VR studies more coherent, our aim is to help better translate empirical findings into practice.

Effectiveness of AR/VR technology depends on many factors such as type of AR or VR technology used, the virtual content provided, the user, and the kind of tasks to be done. Such factors influence how effective AR/VR technologies will be in particular situations. In a wealth of good information, we found clarity issues such as: inconsistent use of terms across research studies (e.g., AR in one study is mixed reality (MR) in another); identification of a system without specifying the features used; and limited description of the tasks a user is expected to perform. Based on study reviews and discussion sessions with AR/VR professionals, we developed and implemented a practitioner-oriented framework to help systematize and bring clarity to the often discrepant research findings.

A. Background

The continuum of MR is a frequently used organizing schema for AR and VR technology systems. Milgram and Kishino (1994) initially described the MR continuum in reference to visual displays. One of the early implementations presented synthetic three-dimensional (3D) visual information with a head-mounted display (HMD) that changed according to the wearer's head movements (Sutherland 1968). AR and VR studies can be placed in the MR continuum, which spans from completely real environments to completely virtual ones, though this is a conceptual continuum versus a continuum with a quantifiable metric.

With AR technology, most of what the user perceives is the real world with some virtual entities mixed in. An example is the heads-up displays (HUDs) for fighter pilots who can see the world around them but information like altitude, heading, and speed are virtually displayed on their visor (Azuma 1997; Azuma et al. 2001). With VR technology, most of what a user perceives is virtual such as in vehicle and flight simulators. For example, students participating in the Air Force's Pilot Training Next program can sit in a chair, strap on a high-resolution headset with 3D spatial audio, and practice maneuvers (Oprihory 2020). VR headsets give the Air Force the opportunity to provide airmen with much more access to simulation technology. We may argue, of course, that AR and VR never occur in a pure form but that would miss an essential point. We can augment and represent reality using many different technologies having complexities that are difficult to organize and understand.

1. AR/VR, Training and Performance Enhancement

Prior analytic reviews of AR/VR depended on two primary tools: meta-analyses and narrative literature reviews. Narrative literature reviews focus on particular topics of interest to their authors as a subjective way to organize research. They provide multiple perspectives on how to think about AR/VR. Meta-analyses are a systematic way to combine data from multiple studies and can identify common effects or identify the reasons for variation. Meta-analyses are optimal when multiple studies aim to answer a specific question or hypothesis. Neither narrative literature reviews nor meta-analyses fully describe dimensions that influence the use and generalizability of AR/VR effectiveness but meta-analyses, in particular, helped to guide the current work.

A factor in understanding AR/VR training effectiveness is that the scientists who do the research come from many specializations. The result is uneven descriptions and findings from one study to another due to the variety of perspectives and their individual, focused goals. We found in an initial review of the literature that somewhere between a quarter and a third of the studies lacked details such as descriptions of users' task experience, AR/VR technology experience, and/or performance measures. We also found inconsistent use of the terms AR and VR themselves and the labeling of technology characteristics, identification of a system without specifying features used, and limited descriptions of tasks trained. To summarize, both the available types of reviews—literature and meta-analyses—and varied details provided by specific studies, prompt the need for an organizing schema to better capture the effectiveness of AR/VR.

2. Factors to Consider in AR/VR Effectiveness

While AR/VR can have value for training, it is important to consider what factors and circumstances influence success. Fletcher et al. (2017) conducted a meta-analysis of 22 AR and augmented virtuality (AV)[3] training effectiveness studies about education, training, and performance aiding. A majority of the analyzed reports of AR/VR compared to alternatives indicate that it saves training time, increases amount learned and its persistence, and results in fewer errors. In addition, users reported that they preferred AR/VR to a number of alternate approaches and that they were more engaged in their tasks.

Specifically, Fletcher et al. (2017) found most reports favor AR/VR effectiveness to alternatives based on effect sizes:[4] AR/VR reduces time (effect size, g = 0.52) and errors (g = 0.81) in performing skilled tasks, increases the amount learned (g = 0.44), produces learning that is more resistant to decay (g = 0.71), is preferred to other approaches (g = 0.81), and increases immersion or *flow* during learning (g = 0.67). Such analyses at the time (i.e., publication dates up to 2015) were comparatively rare.

Recommendations for next steps to advance the value of AR/VR for training effectiveness are many and varied (Fletcher et al. 2017). We need increased emphasis on empirical assessment of AR/VR compared to other systems; research on the development of systems using machine intelligence combined with AR/VR; research on adapting AR/VR to individual differences; exploiting the potential of AR/VR to enhance success in performing military operations in combat service support (e.g., maintenance and repair); cost-effectiveness assessment of AR/VR in both training and performance aiding; and continuing review of costs and advances in commercial AR/VR technology. In addition, there always is the enduring problem of how to minimize potential negative effects on the human/user, such as simulator sickness.

More focused studies allow us to identify what factors are important and why. For example, Kaplan et al. (2021) conducted a meta-analysis of extended reality (XR; a broad term that includes AR, VR, and MR) technologies. Their analysis examined the effectiveness of XR on training compared to traditional methods. They found that XR is about as effective as existing traditional methods that include simulators and other advanced training methods and technology. They reported that effectiveness depends on the tasks and the population being trained. In addition, Kaplan et al. note that the tasks and the populations being trained are too disparate to determine precisely which factors contribute to better training transfer from AR/VR. In general, they suggest that XR's primary value is in providing training where traditional methods may be dangerous or costly. We need to determine more about the factors that affect AR/VR training.

A meta-analysis about what influences the effectiveness of AR on learning gains (Garzón and Acevedo 2019) studied the role of moderating variables. They analyzed 64 quantitative research papers (N = 4705) published between 2010 and 2018 in major journals. Results were that education level and domain of study were two moderating factors influencing the effectiveness of using AR. In addition, Garzón and Acevedo grouped the control treatment (i.e., what the AR

[3] Reference uses AV as a similar term to VR, which is the consistent term in this report.

[4] Hedges & Olkin (1985) g was used to calculate effect sizes.

training system was compared to) in an attempt to see if this had an influence on the effect size of using AR. Their paper (p. 246) "classified the control treatments into three pedagogical strategies: 1) *Multimedia* that refers to educational resources that use different content forms such as videos, images, animation, and learning objects, 2) *Traditional Lectures* that refers to curriculum-based and lecture-based teaching, and 3) *Traditional Pedagogical Tools* that refers to traditional educational resources that teachers use to complement their lectures." The analysis of control treatments suggests that AR has a greater impact on the learning gains of students than these other types of strategies. Overall, it seems that a better understanding of the interaction of moderating variables on training effectiveness may lead to a more complete depiction of when and how AR/VR technology could best be used.

Another aspect of how AR/VR technology may be more or less effective is how it is incorporated into a class or curriculum. A meta-analysis of how AR technology could be incorporated into training methods found a medium-level (effect size, g = 0.637) benefit of AR from 45 studies between 2013 and 2019 (Batdi and Talan 2019). As one example, they report a positive contribution of teaching non-observable subjects (e.g., 3D representations of molecular structures); reducing the rate of error with repetition; and providing innovative, realistic, collaborative, and interactive environments. Similarly, Garzón et al. (2020) in a meta-analysis of AR training interventions found that they have more influence on learning outcomes than pedagogical approaches such as the use of multimedia resources and traditional lectures. The domain to be trained may also influence the likelihood that AR/VR training instantiations will be more or less effective. Some examples include, meta-analyses of using AR/VR technology for students learning scientific concepts (Yilmaz and Batdi 2021); physiology and anatomy (Moro et al. 2021), and surgery (Haque and Srinivasan 2006). There are many subtleties to determining AR/VR technology's relevance to a class or curriculum.

While much of what people focus on for AR/VR technology is the visual display, there are other features in the technology that may have training benefit. For example, with many AR/VR systems there is tracking of the user. This includes where the user is looking (e.g., head orientation, eye tracking), what they are doing with their hands, or where they are positioned in an environment. Limbu et al. (2018) conducted a systematic literature review and analysis of 78 studies that have implemented AR and sensor technology for capturing expert performance to train apprentices. They showed how a range of sensors available to record body and hand movement, force applied, and attentional focus of the user can support an AR methodology for learning environments. Tracking information like this may be useful for both developing training by understanding how an expert might behave in a situation and how to teach it to a novice. As the few examples above illustrate, an understanding of AR/VR effectiveness depends on identifying many factors.

As these few examples illustrate, an understanding of AR/VR effectiveness depends on identifying many factors. The result of our initial look at effectiveness factors leaves as many questions as answers: What are the specific technology characteristics and learning tasks? What are appropriate measures of effectiveness? In what ways do the technology-assisted tasks take less time? How does learning acquisition and its longevity vary by learning modality (e.g., visual, cognitive, auditory, motor, sensorimotor coordination)? AR/VR is an exceptionally diverse research field that demands better organization than existing reviews provide.

B. Rationale for Current Work

Our examination of AR/VR studies shows that the factors used to determine effectiveness are not being identified regularly and need better categorization to provide structure to our understanding of AR/VR training effectiveness. The field requires a more detailed and nuanced organization in order to better facilitate lessons learned, best practices, and identifying gaps.

Part of organizing the field is to promote a taxonomy or framework in which components are organized into groups or types based on particular characteristics. In developing a framework, there needs to be a balance in determining salient categories versus including superfluous categories which may make the framework needlessly complex and unwieldly to use. The AR/VR framework emerging from our analyses include: a) the type of technology, b) the task to be trained, c) who the trainee is, d) the training use or integration into a situation/course, and e) its outcome. The development of this framework described next supports a systematic method and format allowing training developers, designers, and other users to access and understand information they need.

2. AR/VR Framework Methods and Procedures

Development of the AR/VR effectiveness framework consisted of three parts: 1) analyzing prior research efforts on published training effectiveness studies, 2) conducting an initial literature assessment that spanned diverse fields and domains of research, and 3) gathering input from AR/VR stakeholders and practitioners. Prior research provided a starting point for the framework by identifying dimensions of AR/VR study characteristics in areas such as maintenance, operations planning, and observation and control (Fletcher et al. 2017). Results of that work showed that training effectiveness is influenced by task and performance domains (e.g., fast vs. slow movement). An initial review of the literature for this effort consisted of systematically casting a wide net and sifting through approximately 400 AR/VR studies, some that assessed system effectiveness and some that described system developments or implementations. This assessment revealed additional characteristics of interest as a foundation for the AR/VR framework. Discussions with AR/VR stakeholders and professionals confirmed the need for a systematic analysis and organization of AR/VR studies.

A. Components of the Framework

The effectiveness framework features four dimensions and an outcome category represented by specific questions to be answered for each AR/VR study as entries for a knowledge base (discussed in the next section). The dimensions and outcome encompass important study information that provides a comprehensive picture of AR/VR as follows:

1. Technology – AR/VR system and components being assessed: What is the device or system being used?
2. Skill/Task – features of the task/job the user learns: What are the tasks/skills being trained or performed?
3. Training-Technology Integration – features of how the technology is used for training: How is the technology used or integrated in a situation/course? What is the training method?
4. Users – description of the trainee's characteristics, the users of the technology: Who are the performers/trainees? What is the prior skill level of the user for the task and with the technology?
5. Outcome – method of assessment: What metrics were used to measure effectiveness? How effective was the AR/VR in supporting performance?

B. Using the Framework for a Knowledge Base

The AR/VR effectiveness framework aims to compile information from empirical research on training effectiveness. In order to apply the framework to a particular study, two processes take place: 1) Answering the framework questions with "specific" and "general" information, and 2) Coding each category of the framework with appropriate tags.

1. Answering Framework Questions

The answers to each of the framework questions provides both "specific" and "general" detail in narrative form. The "specific" summarizes the instantiation of the study, explicitly describing what in particular was done, how it was done, how it was measured, and its results. The "general" is a description of the characteristics and general implications of the study. For example, the Microsoft HoloLens is a specific technology, while a general description may include the type of device and its characteristics (e.g., a see-through HMD with head-tracking sensors), and how the technology might be employed in the application. The expectation is that the framework information will provide both an understanding of the study's specific outcomes and in general how it contributes to a robust and evidence-driven understanding of AR/VR study characteristics and application.

2. Coding Framework

In addition to the narrative descriptions in the framework, each framework category about a study is tagged with specific codes (see Appendix A for codebook). For example, technology used in a study could be a visual display on a computer monitor or screen (tag = tech_visual computer), and it also can be an occluded HMD device (tag = tech_visual_HMD_occluded). The tags label common characteristics of studies in the knowledge base and can be used during a query to explore gaps (i.e., what tags aren't well populated in the knowledge base), and general characteristics of the research landscape (e.g., intersection of tasks that use occluded HMD technology and have fast motor movements). The minimum is one tag per category of the framework and applying enough more to represent the study.

3. Building the AR/VR Knowledge Base

Using the framework, AR/VR empirical work is assessed and used to populate a Microsoft Access knowledge base.[5] The knowledge base can be queried by targeting specific descriptions (e.g., Microsoft HoloLens; landing an F-14 on a carrier) or more general characteristics (e.g., see-through HMD; eye-hand coordination). The ultimate goal of the knowledge base is to help tease

[5] Microsoft Access is a relational database management system (RDBMS) that is part of the Microsoft Office suite included in the Professional and higher editions or sold separately. Its use in this work should not be interpreted as endorsement of the product.

apart the evidence of AR/VR effectiveness in a more nuanced and systematic way—something that is not done routinely.

The approach of inserting studies into a knowledge base supports many different forms of information, fields, and topics of study. This approach helps to identify clusters of studies for users in three broad fields: technology/system developer, training/instructional designer, and instructors who are domain specialists (e.g., flight, maintenance, or construction).[6] Both the kinds of information and its users are essential input for how to organize the knowledge base.

The result is a multifaceted schema for organizing and coding knowledge base contents. This schema provides a standard set of questions and narratives to characterize information, tags to classify the contents of those narratives, and a search capability to explore varying levels and types of information across studies. The assignment of tags to different kinds of information in studies can guide users toward what may satisfy their interests when assessing AR/VR training effectiveness. This approach allows us to be more systematic in our analysis of the literature and can capture more of the landscape than a simple keyword search. Particular combinations of characteristics (tags) or of phrases can identify clusters that a keyword search would not (e.g., empirical evidence on handheld devices and slow motor movement tasks). Searches across studies with similar characteristics can reveal evidence pointing in the same or different directions. Thus, knowledge base users can select the best ways to describe and find what they need.

4. Identifying and Recording Information in the Knowledge Base

The way to codify the multifaceted schema for organizing and saving information is to develop a template and explanations for data entry (See Appendix B: Analyzing Effectiveness Studies). It encompasses the four framework dimensions and the outcome category. The template describes the information expected in each narrative section, provides space to enter tags, and includes a place for explanatory notes about entries. All of the entries and revisions initially are done with Microsoft Word for ease of entry and editing.

The process for evaluating empirical studies incudes an initial expert-level review followed by a second expert reviewer to validate the information. Initially, each AR/VR study is read by a scientist/practitioner who enters the required information summaries and tags into the template. The validation step includes another scientist/practitioner who reviews and revises the completed template. This process provides more uniform results from one expert reviewer to another. The complied Microsoft Word template is then copied into an Access database that supports search and retrieval of various kinds. The process insures that the template's contents are created and revised separately from entries in the Access database.

[6] No one of these fields is unique, of course, and may be found combined with one another in any specific study.

5. Querying the Knowledge base

The final step in developing the knowledge base was to produce guidelines for assisting users with varying perspectives and interests in finding meaningful output from a query. This step includes determining different ways to organize information and providing instructions with examples. We expect differences in what kinds of information will meet the needs of one kind of user versus another such as an instructional designer or a system developer. For example, an instructional designer may initially be most interested in the effectiveness of an AR/VR method proposed by a study, whereas a system developer may want very detailed descriptions of the technology features and capabilities. When users query the knowledge base, they can select their particular areas of interest for the search. The output of a query provides relevant studies. For each study, the information provided includes a full article reference, a short summary with all components of the framework, the article's abstract, and descriptions of each of the five framework components. To date, we have identified potential users including system developers, program managers, instructional system designers, trainers/instructors, education and training researchers, and training support personnel.

3. Results from the Knowledge Base

The overall result of this effort is an expandable knowledge base, a repository of training effectiveness studies. Currently there are 64 studies in the knowledge base, but there are plans to add studies in the future. The knowledge base functions support multiple search strategies to find empirical evidence most relevant to a user's specific needs. Such searches are satisfied based on the structural and searchable features of the knowledge base—the key to meeting users' needs. We conducted an initial set of searches that address overall effectiveness: technology components, skill/task domains, training-technology integration, and users. The results are presented below.

A. Overall Effectiveness

A knowledge base user's approach to outcomes and their measurement can be from different perspectives. A general question at a high level like, "Are AR/VR systems effective for training?" can be assessed by looking across all of the studies in the knowledge base. In Figure 1, the findings represented with the green bars indicate the count of studies that were categorized as either more effective (22), mixed (15), or less effective (4) than traditional training. In addition, 23 studies were assessments of components of AR/VR systems that investigated the mix of features within a system to determine which ones influence effectiveness. While this is based on just a sampling of studies, it does show that training with AR/VR can be effective, but not equally in all cases.

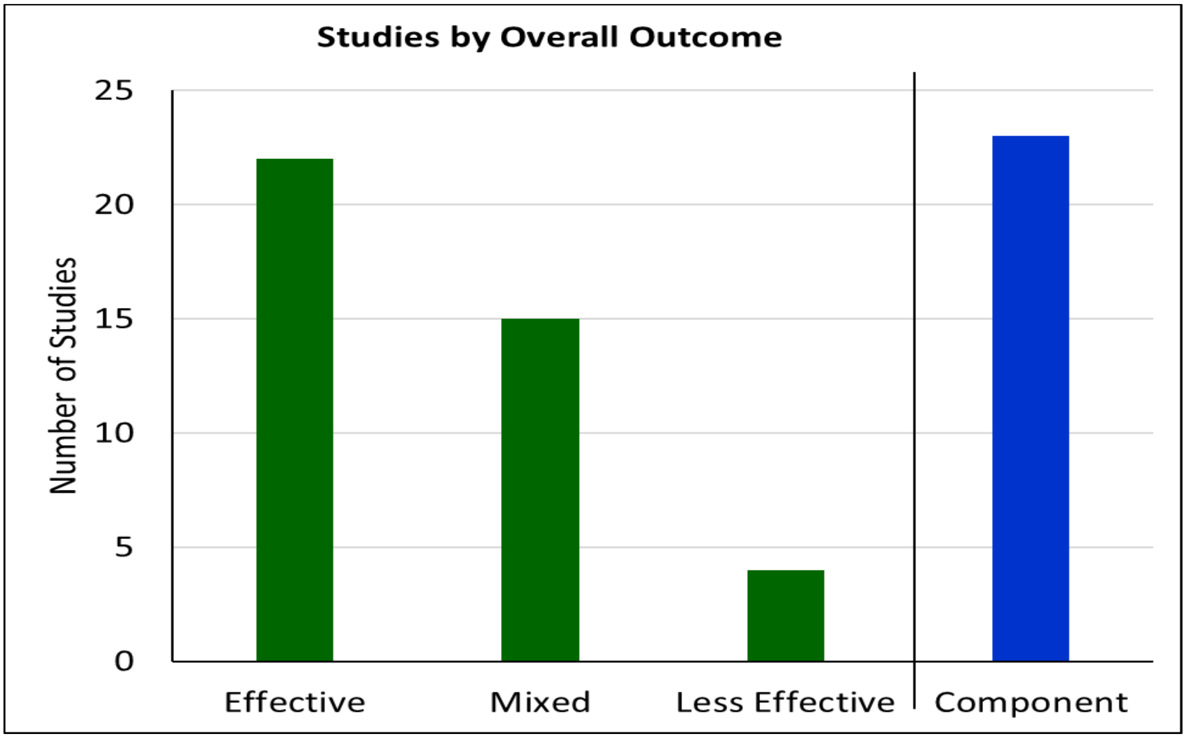


Figure 1. Categorization of studies based on findings about outcomes or system effectiveness: effective, mixed results, or less effective than a traditional system (green bars), or a study that compared components of the AR/VR system (blue bar).

The most common outcome of studies that compared AR/VR to traditional training found that the AR/VR studied was better than a traditional training comparison, followed by those where they were about the same ("mixed"), and a few where AR/VR was less effective. Three medical examples of where AR/VR was more effective for training needle insertion used a projection-based system (Gierwiało et al. 2019), AR glasses (Huang et al. 2018), and a smartphone camera/display (Hecht et al. 2020). An example of the mixed finding (Chalhoub and Ayer 2019) used a see-through HMD for construction layout tasks to reduce large errors (e.g., putting an electrical box on the wrong wall) but was less effective for small errors (e.g., placement of an electrical box inches above the floor). An example of a less effective outcome was Bach et al. (2017) where architecture students using a traditional desktop system for visualizing spatial data performed the analysis tasks faster and with fewer errors than either an AR see-through HMD system or with a handheld tablet AR system. The knowledge base allows its users to focus on which studies are more or less effective.

The other type of study summarized in Figure 1 focuses on refining AR/VR features and functions. An example of such a study is Reiner et al. (2022). Army reservists used an Oculus Rift VR setup with a headset and joysticks to follow a pre-planned path through a virtual city in a reconnaissance task using one of two mapping configurations: (a) mirror in the sky (MitS) where a virtual map was overhead and (b) traditional north-up map positioned as if held in one's hands. Results indicated that the north-up mapping was better than MitS in every metric: subjective preference, accuracy of rerouting, obstacle collisions, participant threat detection, mental workload, and retrospective recall of the path taken. Such studies attempt to identify the relative benefits of particular configurations of the system to refine it and improve overall performance.

1. AR/VR Technology Components

Another possible line of inquiry for this knowledge base is to assess its content for technology components and gaps in them, identifying what technologies have substantial evidence (i.e., used in many studies), and what technologies have little evidence. Table 1 shows the frequency of the most common AR/VR technologies (i.e., hardware) in the knowledge base studies. The different kinds are an indication of the variety encompassed by AR or VR; there are many additional technologies in the knowledge base, but this table shows only the top 10.

The most frequent technology is an occluded HMD, which is in half of the studies. Other visual-display technologies include computer monitors, see-through HMDs, projections on a surface, and handheld tablets. The knowledge base also has non-display technologies such as 31 studies that track a user's hands, 24 that do head tracking, and 12 studies for tracking equipment that a user manipulates or items in the real world that may have markers (e.g., QR codes) for identification. Some of the studies may overlap one another, of course, which a search can identify.

Table 1. The 10 Most Frequent Hardware Technologies Assessed in the Knowledge Base

Technology	N=64
Occluded head-mounted display	32
Tracking user's hands	31
Tracking user's head	24
Computer monitor used for visual display	18
See-through head-mounted visual display	14
Headphones to provide audio to user	15
Tracking system for equipment or mechanical device	12
Tracking system that uses applied markers for registering marked item	8
Projector provides visual information to user	7
Handheld tablet or phone provides visual information to user	7

Not only can the knowledge base query find existing technology evidence, but it also can help identify where gaps in the evidence exist. For example, currently there are only single studies in the knowledge base that describe the tracking of the user's eye/gaze to document where they are looking (Abidi et al. 2019) or the use of haptic gloves to provide feedback about when an object is in their grasp (Cooper et al. 2018). This gap in studies of particular technologies can alert AR/VR researchers to focus their efforts on the need for development while trainers can identify what is not supported by rich empirical evidence.

By categorizing studies by technology, the knowledge base can help a user find studies for an application of interest. Then, a user can probe those studies to gain an understanding of the particulars of that technology, when it may be an effective tool, when it might not, or how it could be paired with other technologies or training characteristics. For example, a training developer interested in the effectiveness of an occluded HMD would search for and see summaries of all 32 studies in the knowledge base (see Figure 2). Of those 32 HMD studies, 9 found AR/VR to be more effective than traditional training, 6 had mixed results, 2 showed traditional training to be more effective, and 15 provided a component comparison for the features of the system that might lead to more or less effectiveness.

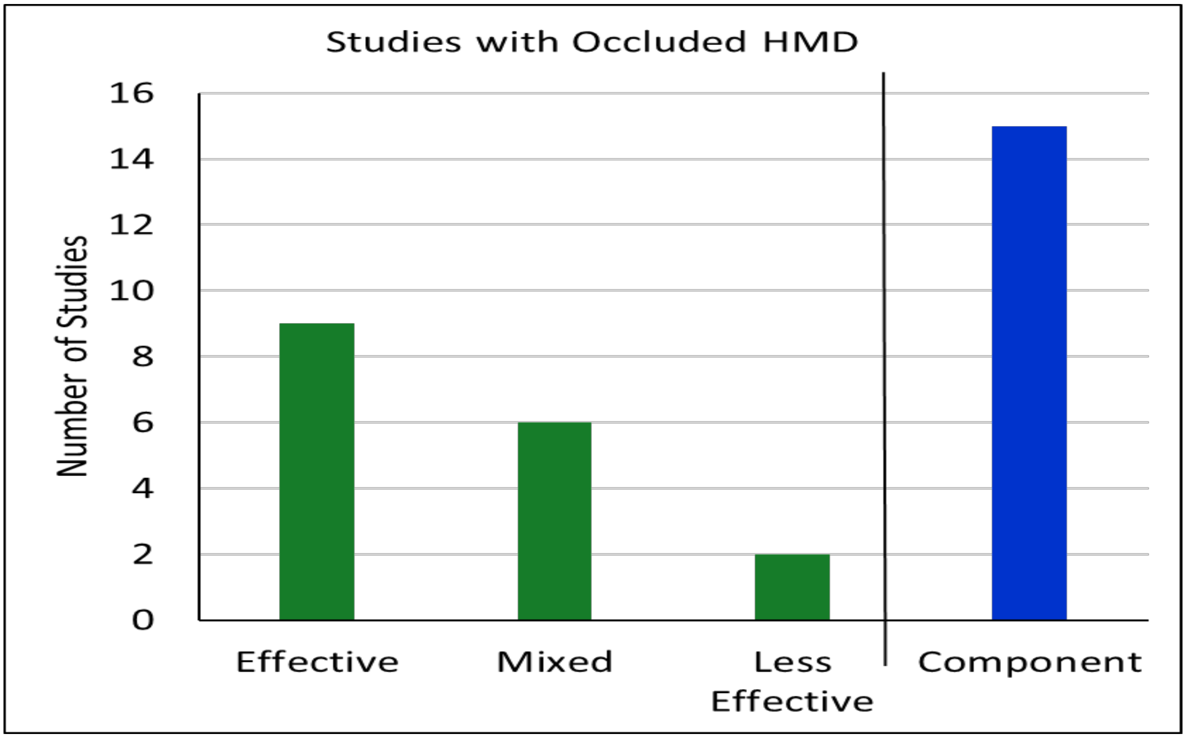


Figure 2. Studies that include occluded HMD grouped by outcomes: effective; mixed results, less effective than a traditional system (green bars), or the study compared components of the AR/VR systems (blue bar).

By going through the summaries, the training developer can learn about the different ways that such a device has been used effectively such as: coordinating aerial firefighting helicopters (Clifford et al. 2020); improving performance of novice robot teleoperators (Brizzi et al. 2018); increasing safety for crane operators (Dhalmahapatra et al. 2021); and increasing students' ability to recognize and correct problems while parachuting (Liang et al. 2020). This query also provides cases where the VR training was less effective than the traditional training method. For example, Winther et al. (2020) found that using an occluded HMD (HTC Vive) for maintenance of a pump part was not as good as actually holding parts and manipulating them; Oberhauser et al. (2018) found that licensed pilots practicing flight maneuvers performed poorer on a VR simulator than the traditional hardware simulator. They also rated the VR simulator as having higher workload and causing more simulator sickness. Reading such summaries can provide the knowledge base user with ideas on problems to avoid.

Additionally, there are studies that assessed the subcomponents of a system to show which system configurations worked relatively better or worse. In one such study, Westerfield et al. (2013) demonstrated how the addition of an intelligent feedback system improves the training of assembly tasks; another study, Monteiro et al. (2020), showed the importance of different sensory cues in firefighting training. Knowledge base searches of any of the framework components can provide a training developer with ideas about how the searched-for category can be used and under what conditions it might be effective.

Another frequently studied technology is hand tracking with 31 studies incorporating it into AR/VR training. For those studies, Figure 3 shows 10 where AR/VR was more effective than traditional training, 6 with mixed results, and 3 in which traditional training was more effective. There were also 12 studies that provided a component comparison for the features of the system that might lead to more or less effectiveness.

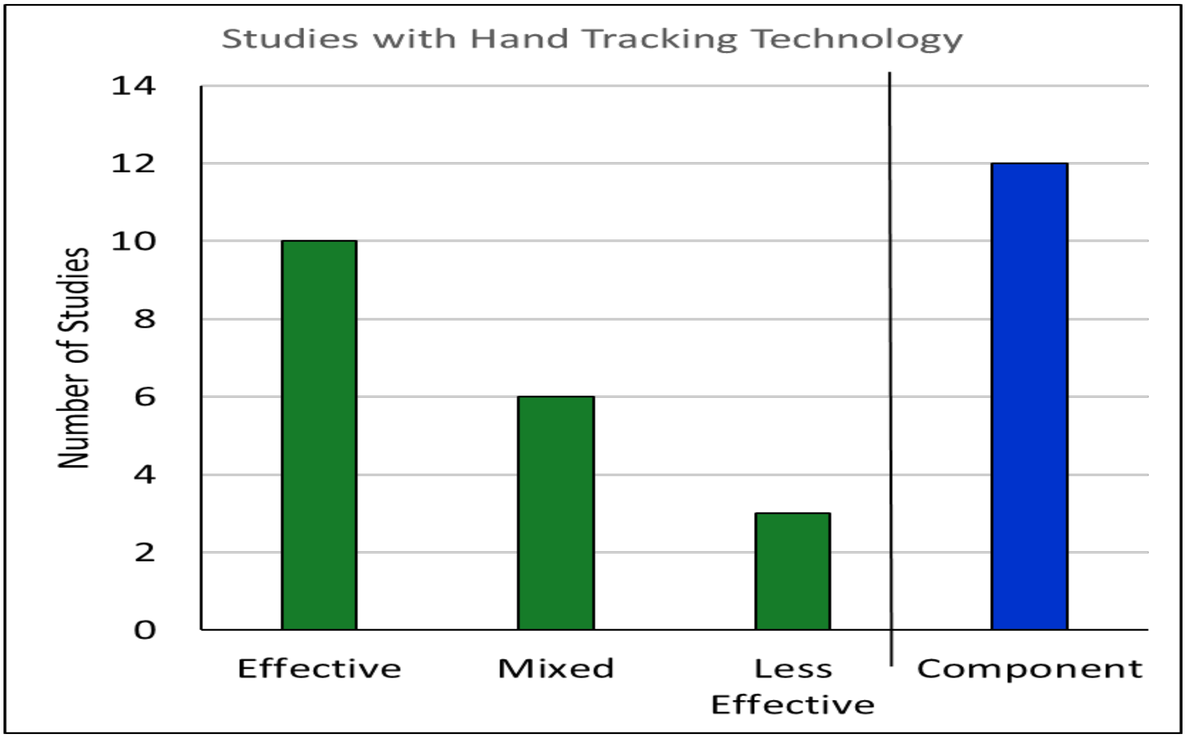


Figure 3. Studies that include hand-tracking technology grouped by outcomes: effective, mixed results, or less effective than a traditional system (green bars). The blue bar represents studies that compare components of AR/VR systems.

By searching for and reviewing summaries incorporating hand-tracking technology, the training developer would learn about the different ways it has been used effectively such as training military medical skills (Siu et al. 2016), parachuting skills (Liang et al. 2020), and manufacturing/assembly skills (Abidi et al. 2019). This query would also provide cases where the results were mixed such as flight training (Fussell and Hight 2021) and aircraft maintenance (Horner et al. 2020) as well as where training was less effective than the traditional training method (e.g., assessment of 3D visualization and analysis of data, Bach et al. 2017). Additionally, there are component studies including Ragan et al. (2015) which showed how field of view and scene complexity influenced VR scanning behavior when users had to detect and track potential threats in a virtual city environment; Bailey and Johnson (2020) who assessed how natural hand gestures can be used to interact with the system; and Frederiksen et al. (2020) who found that different levels of immersion influenced time to complete tasks and hand-movement efficiency in laparoscopic surgery.

2. Training Domains

Within the knowledge base, there are studies that represent many different training domains and skills. Discussions with potential users indicated that their first knowledge base query might be for domains (i.e., skills and task) like ones they were planning to train. Figure 4 shows the four most frequently trained domains that queries found. The most common is maintenance and assembly tasks with 19 studies, followed by 16 medical studies, 8 for flight training, and 8 about military tasks. For both maintenance and medical domains, the most likely outcome was that AR/VR was more effective than traditional training. For flight training, the most frequent outcome was mixed, but a caveat to this finding is that several studies compared the new AR/VR system to a well-established flight simulator. For the military tasks (e.g., aiming and firing a rifle, conducting a reconnaissance or rescue mission, planning a military operation), the most frequent were component evaluations to determine system refinements based on how components influenced effectiveness, indicating that these applications may still be in a development stage to see how they can be made most effective. Other domains than those summarized here had much lower frequency or were contrived tasks (e.g., moving a box through a maze) that fit a lab/experiment setting but not a clearly named domain.

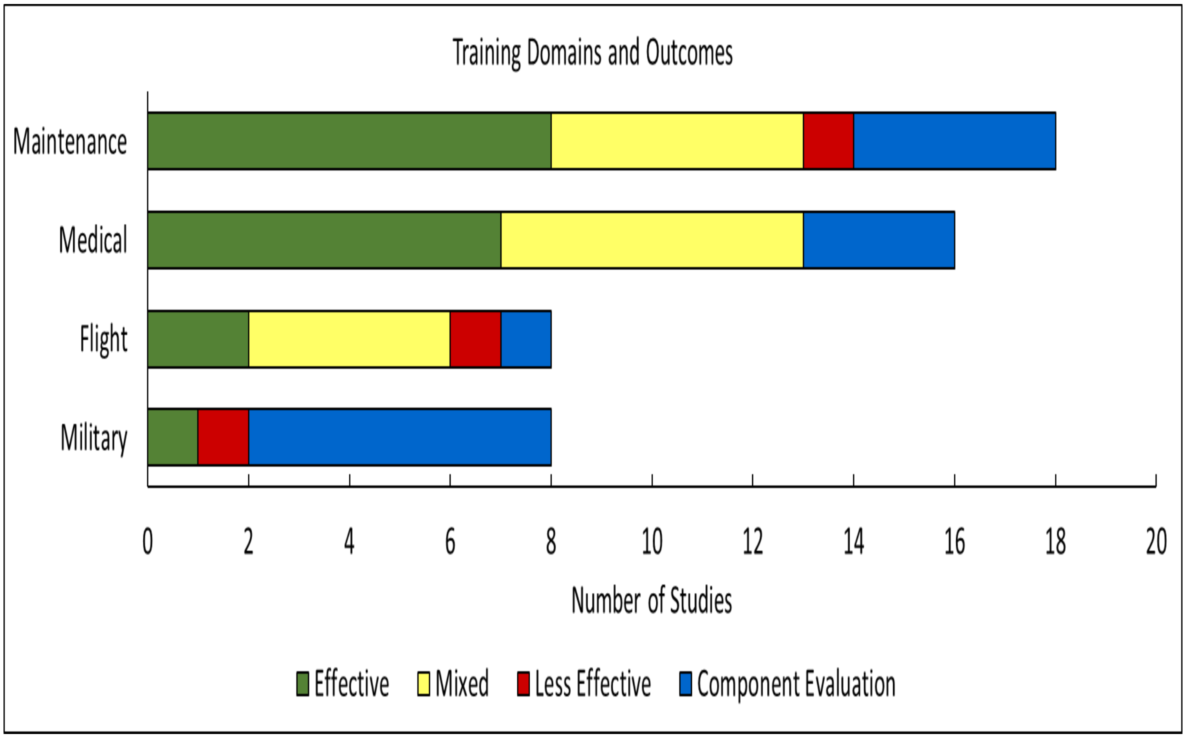


Figure 4. The frequency of different training domains in the knowledge base color coded to represent when the studies' outcomes were effective, mixed, less effective, or were component evaluations.

In addition to the domain topics, the skills/tasks can be categorized by their physical characteristics. These include tasks that require a user to coordinate what they see with how they move their hands such as surgical knot tying (Yoganathan et al. 2018), grasping and manipulating an object with a teleoperated robot (Brizzi et al. 2018) or employing some combination of physical movements including firing a rifle (Bhagat et al. 2016; Pettijohn et al. 2019), changing the alternator in an airplane (Bailey et al. 2017), or conducting maintenance on an industrial pump (Winther et al. 2020). Below in Table 2 are some of the generic descriptors that characterize the different types of skills/tasks trained with AR/VR technology. These descriptors are not exclusive in that a study might include training that can be characterized in multiple ways such as some surgical skills that require eye-motor coordination as well as a fine motor skill.

Table 2. List of Different Types of Skills/Task Descriptors and Their Frequency in the Knowledge Base

Skill Type	Study Count = 64
Eye-motor coordination	40
Visually discriminate	33
Acquiring knowledge	27
Fine motor skill - manipulation	22

Skill Type	Study Count = 64
Decision making	20
Slow motor action	19
Fast motor action	10
Auditory detection/orientation	9
Touch-motor coordination	7
Gross motor skill	6
Visually tracking movement	3

The intent of this analysis by generic skills/task descriptors is to point out those studies that may be relevant to a user's interest or that share some characteristics with that interest. For example, some fine motor surgical skills may share characteristics with fine motor skills like assembling small equipment. Knowledge base searches can help identify lessons learned at intersections of disparate domains because of the similarities in the fundamental actions. Two studies using haptic feedback illustrate the value of such knowledge base intersections. In one (Li et al. 2019), urologists used a VR device including fine motor (haptic) feedback to perform a simulated biopsy. Novice urologists preferred the VR system more so than experienced ones in guiding the biopsy. In another study (Winther et al. 2020), VR used such fine motor feedback for training the maintenance of an industrial pump. It helped VR to be effective, but actually holding parts and manipulating them is even better. The two studies considered together and with others as well can help inform research about related findings in technologies and techniques.

This level of distinction can also be helpful when a higher-level description encompasses many different types of actions that a query needs to separate as with the set of military tasks in the knowledge base. For example, the military domain conflates skills such as eye-hand coordination for aiming and firing a gun (Bhagat et al. 2016; Pettijohn et al. 2019), strategic decision making and visual discrimination for a battlefield (Hale et al. 2019), acquiring knowledge and decision making when calling in close air support (Marraffino et al. 2019), and visual discrimination for detecting an improvised explosive device (Reed et al. 2019). Generic queries about tasks/skills can help the user tease apart complex actions into their components.

3. Training Use

How an AR/VR system is incorporated into a training situation is another way that the knowledge base can differentiate between studies. As shown in Figure 5, the most frequent was using the AR/VR system to practice the training task with 42 such studies. These systems let the user do the training task over and over to refine their skills, and may or may not include some form of added feedback or instruction. The next most frequent training use was the overlaying of extra symbols or information on the visual display to provide cues or information to the user that normally is not available without AR/VR (30 studies). Two different types of feedback delivery were studied, with the more frequent being some form of automated feedback about the user's performance level so the user could work to improve (17 studies). Instructor feedback occurred in 12 studies where a live instructor observed the training, evaluated performance of the user, and then provided some form of instructional feedback. While debatable if it is a training application, the use of the AR/VR systems as a job aid (12 studies), might require less initial training to mastery if it is available post-training.

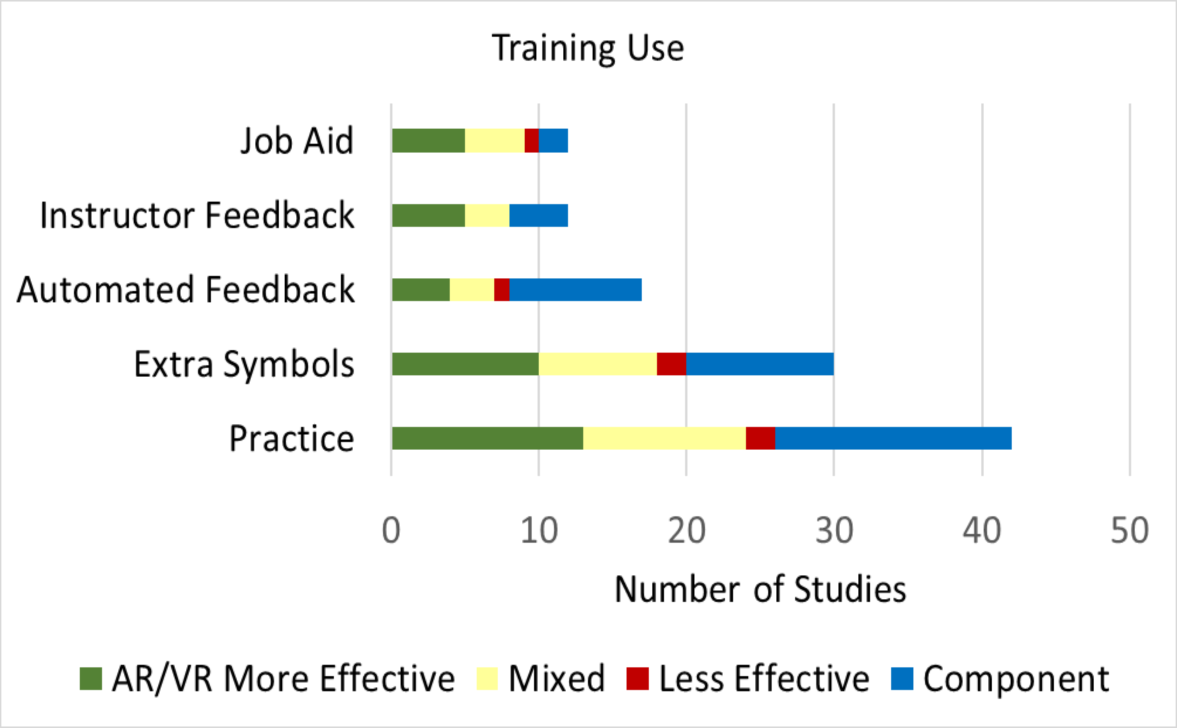


Figure 5. The different ways that AR/VR technologies can be incorporated into a training situation.

There is no clear best way to integrate AR/VR technology into training because the effectiveness of each use has been shown in different kinds of studies. For example, effective use of AR/VR for practicing tasks include students doing parachuting without risking injury (Liang et al. 2020); overhead crane operations to improve safety in construction (Dhalmahapatra et al. 2021); and inserting a nasogastric tube without potentially injuring a patient (Aebersold et al. 2018). There also are effective systems that include extra symbols and information to train users such as extra flight information on pilots' smart glasses to assist them in a challenging flight simulation (Haiduk 2017); directions and progress feedback overlaid onto the real world as users assemble a robot (AlNajdi et al. 2018); and the visual overlay of a syringe on a phantom liver to guide medical students into a specified spot (Gierwiało et al. 2019). Examples of effective use of automated feedback include: Song et al. (2021), where students learned construction crane operations and Abidi et al. (2019), where users assembled street scooters. Examples of effective use of AR/VR with instructor feedback include: Army aviator students learning to pilot a helicopter with a live instructor providing feedback on their performance (Dalladaku et al. 2020), and Bhagat et al. (2016), where students improved their rifle marksmanship on stationary and moving targets. Examples of effective use of AR/VR as a job aid include: Hecht et al. (2020), where medical students inserted a needle at the correct entry point at the right angle and to an appropriate depth, and Kwiatek et al. (2019), where engineering students learned to construct a metal pipe assembly. The varied training uses emphasize the importance of exploring studies to understand the importance of what works and why, which is in agreement with Garzón et al. (2020).

4. User Characteristics

Understanding the initial state of the user (i.e., trainee) prior to the training is important because experience with virtual environments and related technology has been shown to influence subsequent performance (Smith and Du'Mont 2009; Orvis et al. 2005). In the current knowledge base, there are 38 of 64 studies that describe the user's experience level with the technology or the training task. Of the studies that provided information to assess trainees' prior experience level, Figure 6 shows the number of studies and their types of findings (i.e., comparison to traditional method or comparison within an AR/VR system). They are grouped by the user's characteristics such as novice or experienced with either the training task or with AR/VR technology. The blue bars show that many studies included participants who are novices with the technology and also a task novice when compared to traditional methods. Likewise, the orange bars indicate that many studies included participants with prior task experience who were novices with the technology. To a lesser frequency, studies included participants who were task novice and tech experienced (gray bars) or task and tech experienced (gold bars).

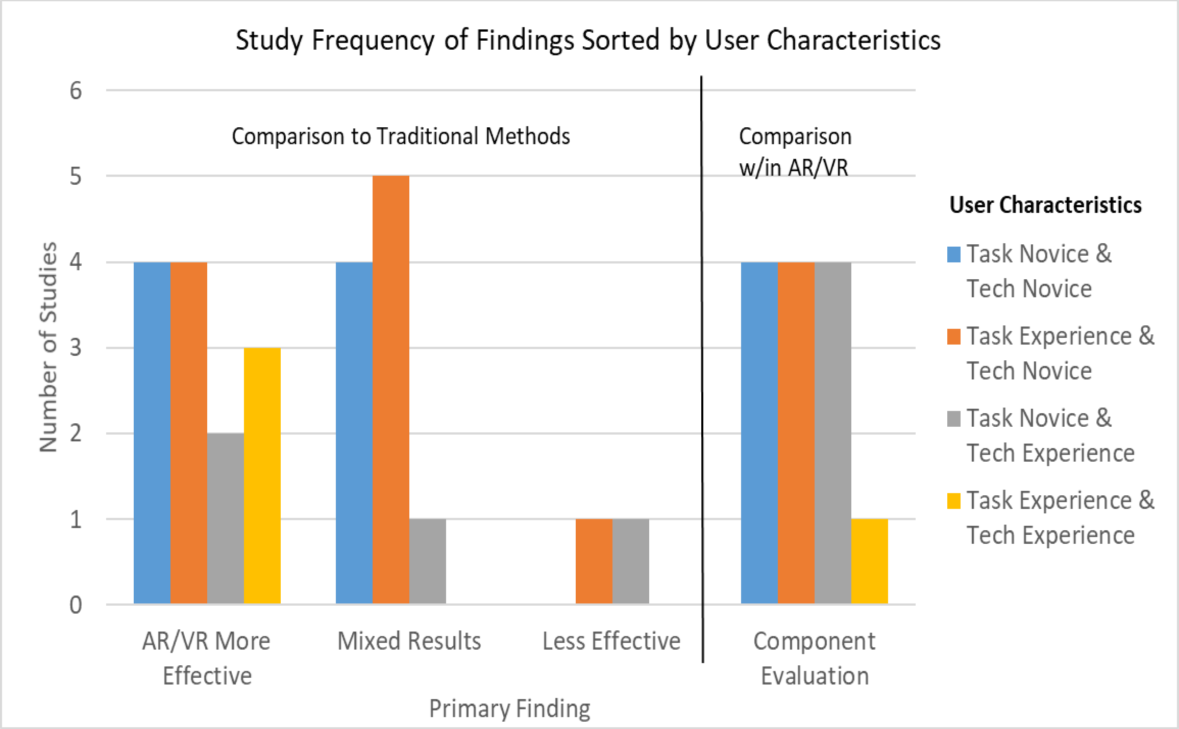


Figure 6. The number of studies with users grouped by their level of prior task and technology experience.

The level of user experience for the task to be trained is an important consideration in understanding if a study is well-structured and relevant to a particular training context. Without a description of the participants, it is not clear what population the study results may or may not generalize to. There is no definitive best type of participant experience level; it is dependent upon how the AR/VR may be incorporated into a course or plan of action. There are situations where you would expect a novice trainee (i.e., initial experience in an area) to support their use. For example, initial medical training provides a good example where students learn how to insert a needle into a manikin's vein while using AR glasses displaying instructional images and step-by-step procedures (Huang et al. 2018). Conversely, there are situations where studies with experienced participants may be appropriate. For example, the test of an AR tablet as a job aid used experienced pipefitters to assemble and inspect a piping configuration in a construction project (Kwiatek et al. 2019).

Experience may influence the outcomes of effectiveness studies, as demonstrated when those both with and without relevant experience were assessed in the same study. For a task experience example, Fan and Wen (2019) showed that those with prior live training in a military task performed better in the VR environment than those without such pre-training experience. For a technology experience example, Brizzi et al. (2018) used participants both with and without prior VR experience and found that on average those with prior technology experience performed the training task (i.e., controlling a robot) faster than those without prior VR experience. The person searching the knowledge base needs to be sensitive to the importance of experience in the value of AR/VR studies

5. Assessment Metrics

A knowledge base user also may approach outcomes and their metrics from different perspectives. Six clusters of metric types that appear frequently are characterized as follows:

- Physical performance such as how well a person completed a task without errors
- Ratings of performance (e.g., Likert scale) that show, for example, how well system users perceive a task was accomplished
- Cognitive performance (i.e., how well a person knows information) as indicated by measures such as a multiple-choice knowledge test
- Usability or assessing how useable a system is for completing a task (e.g., System Usability Scale)
- Workload as a measure (e.g., National Aeronautical and Space Administration (NASA) task load index (TLX)) of the perceived effort that system users require for a task
- Opinion where a system's users judge its utility for performing a task

Studies often use more than one type of metric to provide a better perspective of a system's effectiveness, e.g., physical performance to show how well a skill was learned combined with measuring usability or opinion. Figure 7 shows the frequency of different metrics and their effectiveness in the knowledge base. Overall, for every type of metric, AR/VR is usually more effective than traditional methods or, at least, the results are mixed, with only a few studies indicating that AR/VR is less effective.

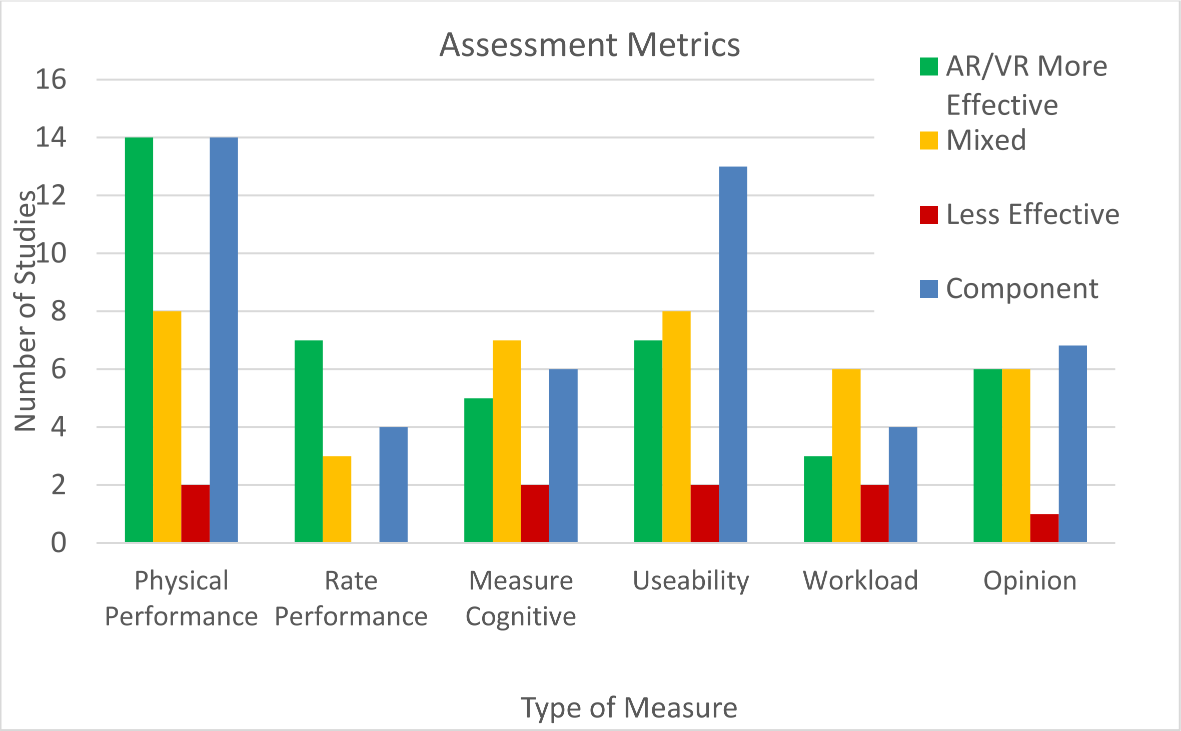


Figure 7. The frequency of different assessment metrics in the knowledge base color coded to represent when their use provided effective, mixed results, or less effective outcomes for AR/VR than traditional methods.

Probing results for assessment metrics further shows that 38 of 64 studies include objective measures (e.g., time, error rate) of how well the users physically performed the training task. Of those 38 studies, 14 demonstrated that training with the AR or VR system improved performance more than a traditional training method, 8 indicated mixed results, and only 2 studies showed that AR/VR training was less effective than the traditional methods. An example of a study where physical performance was measured and the system was demonstrated to be more effective than traditional training is Hou et al. (2015), where the overlaying of instructional images over live video feed acted as a visual aid and allowed assemblers to build a pipe structure in a construction setting faster and with fewer errors.

When physical performance was rated, the results were relatively similar with the highest frequency being studies that show AR/VR to be more effective than traditional methods, followed by mixed results; no studies indicated the system was less effective. Liang et al. (2020) provides an example of a study where users rated physical performance and the system was more effective than traditional training. In that study, military participants rated themselves as more capable following use of a VR parachute training simulator (occluded HMD; head, hand, and body tracking) for dealing with obstacles and malfunctions.

When cognitive measures were used, the most frequent finding was mixed results, followed by AR/VR being more effective, and only two studies where AR/VR was less effective. An example of a study where cognitive outcomes were measured and the system was demonstrated to be more effective than traditional training is AlNajdi et al. (2018). Students used an AR tablet to assemble and learn about a mobile robot in a comparison with students who used a paper-based approach. The AR students had superior knowledge learning outcomes as demonstrated by 10 multiple-choice questions assessing knowledge/understanding of hardware and software components.

Usability and workload are important measures in that they indicate how well a potential user may employ a system to perform the expected task. If a training system has high usability, people are more likely to decide to use the system when given a choice and may be more effective at using the system to achieve a training goal. For example, Corelli et al. (2020) did a study where usability provided the clearest distinction among variations of a VR training system for firefighters while performance measures across systems did not differ much. Regarding workload, Chalhoub and Ayer (2019) had participants rate cognitive workload in a study that compared an AR system (see-through HMD as a job aid) to a traditional method (paper plans) for installing electrical outlets at a building construction site. He found that workload was significantly lower using the AR system while results for other performance measures were mixed (i.e., reduction of big errors but more small errors).

There also are qualitative methods for assessing outcomes, like the opinions of experts or system users who describe how successful a system might be for accomplishing a particular task. Qualitative measures can provide nuanced details as to why something may be more or less likely to work or situations and context that might influence subsequent system designs or implementations. An example of how opinions can be used in this way is Reed et al. (2019), where the Army seeks to identify cost-effective methods to deliver mine detection training and developed two potential XR systems to compare with the traditional method. While the results of the system comparisons were mixed, the opinions provided recommendations for improvements to the technology, including more realistic VR scenarios and updated software.

6. Intersection of Framework Components

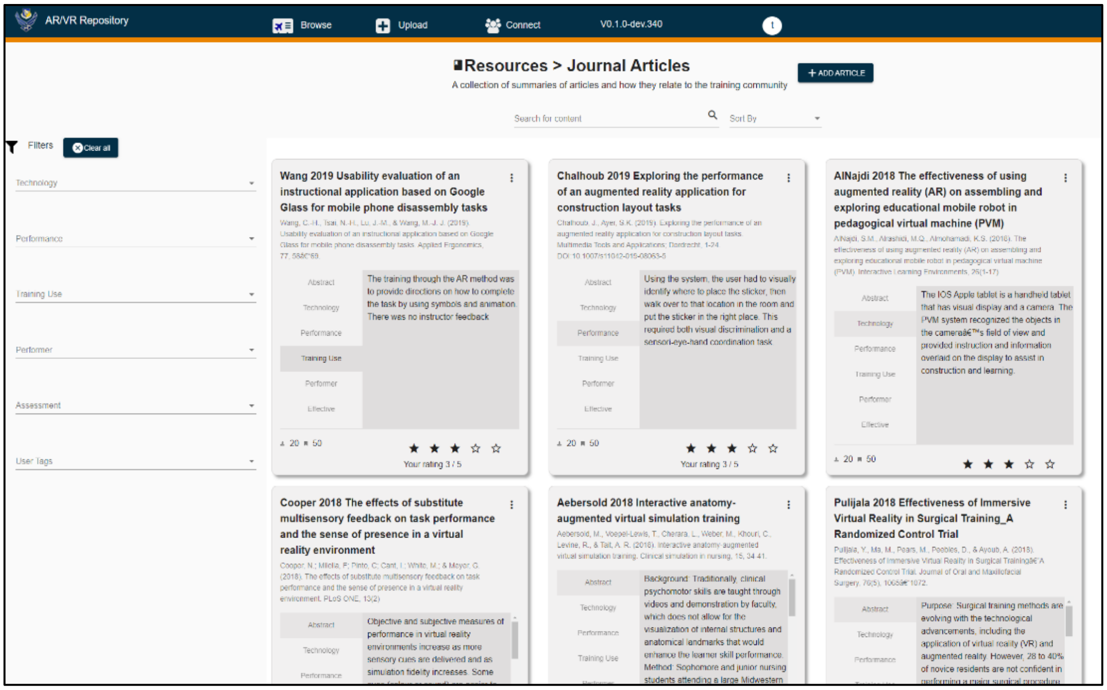
A powerful benefit of knowledge base queries is that they can cut across framework components that are specified by a user. For example, a knowledge base user can not only see the range of technologies included in AR/VR training effectiveness studies, but also can drill down deeper into how a particular technology might be used with a specific training method as part of a training system. To illustrate, the knowledge base returns eight studies for the query: “How can occluded head mounted visual displays [i.e., a particular technology] be incorporated into training with automated feedback [i.e., a particular training method]?” Two of those studies compared the VR systems to traditional training with different results: one found that an immersive VR system outperformed conventional training on two behavioral transfer tests (Makransky et al. 2019); the other demonstrated that a developmental version of VR training was effective but not as effective as the well-developed traditional training system (Winther et al. 2020). Four other studies compared different features of VR training to one another. Those findings included: intelligent tutoring capabilities improved training outcomes over VR training without intelligent feedback (Westerfield et al. 2013); the types of feedback from movement/locomotion in virtual training influenced system usability and performance (Corelli et al. 2020); and matching types of sensory stimulation for feedback influenced training effectiveness (Batmaz and Stuerzlinger 2021; Monteiro et al. 2020). Lastly, there were two studies describing the benefits of VR training with automated feedback, such as decreasing training costs and enhancing training safety (Kwegyir-Afful and Kantola 2021; Song et al. 2021). Such findings provide a combination of technology and feedback mechanisms that trainers/training developers can probe to identify relevant findings for their use case.

By combining framework components and searching for nuances, knowledge base users can probe to a level-of-detail up to and including individual studies. They can identify relevant information for their specific needs or determine which AR/VR studies may not be relevant. They can query the knowledge base by combining its many different dimensions in cross-walks of technology type, skill/task trained or performed, training use, users, and outcomes.

4. Usability Test

The practical utility of the framework and resulting knowledge base depend on usability tests to show how well they work and what changes are needed. As part of another effort with similarities to the knowledge base, Aptima (Beaubien et al. 2022) did the first knowledge elicitation, design, development, and usability testing of the knowledge base. This parallel effort is to build a portal where people could go to access assets and information regarding the use of AR/VR for training. Feedback caused changes (e.g., clarification of the different task types: cognitive, motor, psychomotor) used in the knowledge base.

Figure 8 shows a screen capture of the knowledge base search page as it was for usability testing and before the terminology changes in the earlier part of this paper.[7] The knowledge base includes search, sort, and filtering capabilities. Additionally, each page is dynamic. For example, left clicking on the "Abstract" tab provides a high-level summary of the article. Similarly, the "Technology" tab displays the specific technology platform that was used.



Resources > Journal Articles

A collection of summaries of articles and how they relate to the training community

Wang 2019 Usability evaluation of an instructional application based on Google Glass for mobile phone disassembly tasks

Wang, C.-H., Tsai, N.-H., Lu, J.-M., & Wang, M.-J. J. (2019). Usability evaluation of an instructional application based on Google Glass for mobile phone disassembly tasks. Applied Ergonomics, 77, 58â€"69.

The training through the AR method was to provide directions on how to complete the task by using symbols and animation. There was no instructor feedback

Your rating 3 / 5

Chalhoub 2019 Exploring the performance of an augmented reality application for construction layout tasks

Chalhoub, J., Ayer, S.K. (2019). Exploring the performance of an augmented reality application for construction layout tasks. Multimedia Tools and Applications; Dordrecht, 1-24. DOI:10.1007/s11042-019-08063-5

Using the system, the user had to visually identify where to place the sticker, then walk over to that location in the room and put the sticker in the right place. This required both visual discrimination and a sensori-eye-hand coordination task.

Your rating 3 / 5

AlNajdi 2018 The effectiveness of using augmented reality (AR) on assembling and exploring educational mobile robot in pedagogical virtual machine (PVM)

AlNajdi, S.M., Alrashidi, M.Q., Almohamadi, K.S. (2018). The effectiveness of using augmented reality (AR) on assembling and exploring educational mobile robot in pedagogical virtual machine (PVM). Interactive Learning Environments, 25(1-17)

The IOS Apple tablet is a handheld tablet that has visual display and a camera. The PVM system recognized the objects in the cameraâ€™s field of view and provided instruction and information overlaid on the display to assist in construction and learning.

Cooper 2018 The effects of substitute multisensory feedback on task performance and the sense of presence in a virtual reality environment

Cooper, N.; Milella, F; Pinto, C; Cant, I.; White, M.; & Meyer, G. (2018). The effects of substitute multisensory feedback on task performance and the sense of presence in a virtual reality environment. PLoS ONE, 13(2)

Objective and subjective measures of performance in virtual reality environments increase as more sensory cues are delivered and as simulation fidelity increases. Some

Aebersold 2018 Interactive anatomy-augmented virtual simulation training

Aebersold, M., Voepel-Lewis, T., Cherara, L., Weber, M., Khouri, C., Levine, R., & Tait, A. R. (2018). Interactive anatomy-augmented virtual simulation training. Clinical simulation in nursing, 15, 34-41.

Background: Traditionally, clinical psychomotor skills are taught through videos and demonstration by faculty, which does not allow for the visualization of internal structures and anatomical landmarks that would enhance the learner skill performance. Method: Sophomore and junior nursing students attending a large Midwestern

Pulijala 2018 Effectiveness of Immersive Virtual Reality in Surgical Training_A Randomized Control Trial

Pulijala, Y., Ma, M., Pears, M., Peebles, D., & Ayoub, A. (2018). Effectiveness of Immersive Virtual Reality in Surgical Trainingâ€"A Randomized Control Trial. Journal of Oral and Maxillofacial Surgery, 76(5), 1065â€"1072.

Purpose: Surgical training methods are evolving with the technological advancements, including the application of virtual reality (VR) and augmented reality. However, 28 to 40% of novice residents are not confident in

Figure 8. Knowledge base search page as it was for the usability test. It has since changed based on user feedback.

[7] Other labels and descriptions shown in the figure have changed to match those in this report as a result of feedback from the user test as follows: "performance" = "skill/task," "performer" = "user," and "effectiveness" = "outcome."

Clicking on an individual page brings up detailed information about that article, as well as a link to the full article if available. All of the information is stored in self-contained sections, thereby allowing the user to quickly locate the specific information of interest. At the top of the page is a search bar. Any information typed in this search bar is "filtered in" thereby allowing the user to quickly locate the specific search term in the body text. For example, typing the words "Google Glass" would hide all sections that do not contain the term "Google Glass." Finally, the screen includes multiple different "views" of the data. For example, selecting the "System Developer" view prioritizes technology-centric information at the top of the screen, while selecting the "Instructional Designer" view prioritizes training effectiveness information and the "Domain Specialist" view prioritizes task performance-related information. All of the same information appears on the screen; it is just prioritized in different orders.

A. Method for Usability Test

Aptima conducted this system usability test with five candidate end users.[8] The sample of users included a mix of active-duty military personnel (n = 1), Department of Defense (DoD) civilians (n = 1), and contractors (n = 3). The participants' work-related roles included a mix of instructional designers/developers (n = 2), graphic designers (n = 2), game designers (n = 1), research scientists (n = 1), software engineers (n = 1), and domain subject-matter experts (n = 1). Several of the participants performed multiple roles. All of the usability tests were conducted remotely. The median interview length was 60 minutes, with a range of 50–75 minutes. The following descriptions and outcomes were extracted from Aptima's interview notes to focus on the knowledge base only.

B. Analysis and Outcome

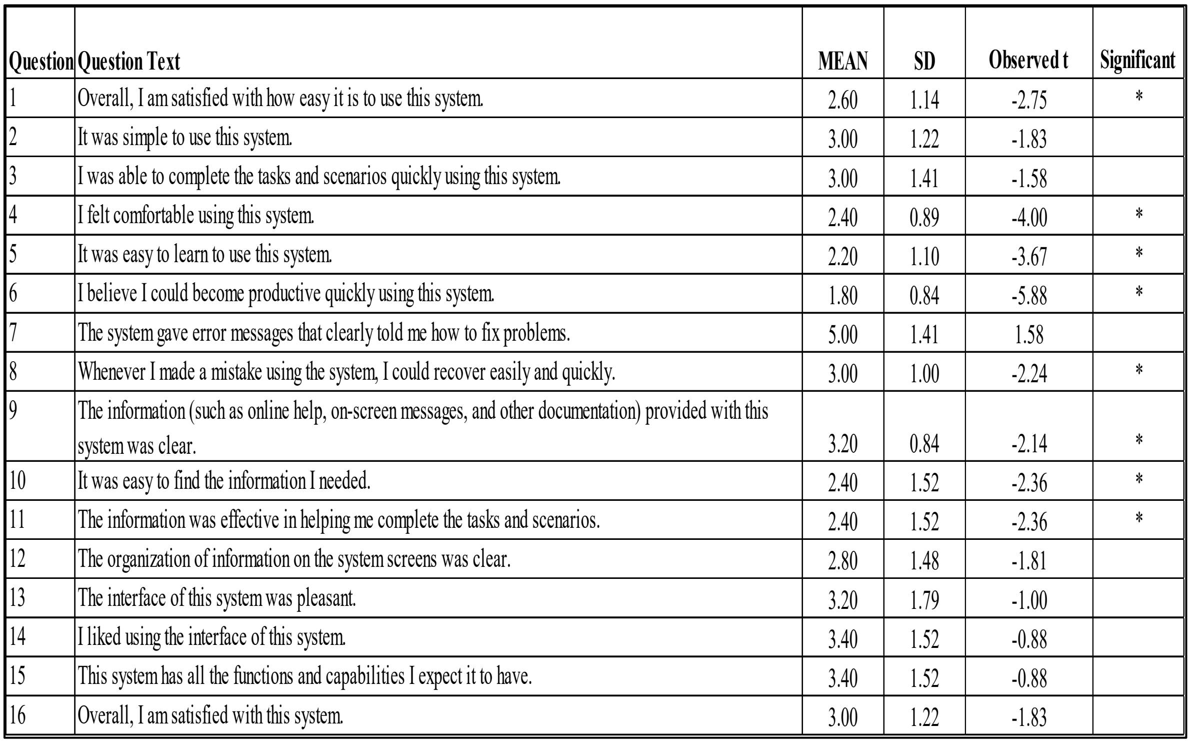
The primary criterion measure was the Post-Study System Usability Questionnaire (PSSUQ). The PSSUQ is a 16-item usability scale that measures perceived satisfaction with a system. The PSSUQ has question anchors that range from "Strongly Agree" (1) to "Strongly Disagree" (7), and has demonstrated high levels of internal consistency reliability (.97) (Lewis 1995). The scale is frequently used in Human-Computer Interaction (HCI) research. The PSSUQ questions are organized into three categories. Questions 1–6 focus on the system/process quality. Example questions include "It was simple to use this system" and "I was able to complete the tasks and scenarios quickly using this system." Questions 7–12 focus on the quality of information provided by the system. Example questions include "Whenever I made a mistake using the system, I could recover easily and quickly" and "It was easy to find the information I needed." Finally, Questions 13–16 focus on the interface quality. Example questions include "I liked using the interface of this system" and "This system has all the functions and capabilities I expect it to have."

[8] Previous research suggests that a sample size of five is sufficient to identify 85% of usability problems (Faulkner 2003).

After providing informed consent to participate and to be recorded, the users were tasked with: 1) searching for and downloading an article from the knowledge base, and 2) uploading and annotating an entry into the knowledge base. As users performed the tasks, they were instructed to talk out loud. After performing a search-download and an upload-annotate task, the participants completed the PSSUQ questionnaire. All of the sessions were digitally recorded for additional analysis.

Mean PSSUQ scores are summarized in Table 3 below. The item means were compared to the scale midpoint (a value of 4), using a one-tailed, single sample t-test. To be considered statistically significant, the mean score had to be smaller than the critical t-value of -2.02.

Table 3. The Results of the PSSUQ



Question	Question Text	MEAN	SD	Observed t	Significant
1	Overall, I am satisfied with how easy it is to use this system.	2.60	1.14	-2.75	*
2	It was simple to use this system.	3.00	1.22	-1.83	
3	I was able to complete the tasks and scenarios quickly using this system.	3.00	1.41	-1.58	
4	I felt comfortable using this system.	2.40	0.89	-4.00	*
5	It was easy to learn to use this system.	2.20	1.10	-3.67	*
6	I believe I could become productive quickly using this system.	1.80	0.84	-5.88	*
7	The system gave error messages that clearly told me how to fix problems.	5.00	1.41	1.58	
8	Whenever I made a mistake using the system, I could recover easily and quickly.	3.00	1.00	-2.24	*
9	The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.	3.20	0.84	-2.14	*
10	It was easy to find the information I needed.	2.40	1.52	-2.36	*
11	The information was effective in helping me complete the tasks and scenarios.	2.40	1.52	-2.36	*
12	The organization of information on the system screens was clear.	2.80	1.48	-1.81	
13	The interface of this system was pleasant.	3.20	1.79	-1.00	
14	I liked using the interface of this system.	3.40	1.52	-0.88	
15	This system has all the functions and capabilities I expect it to have.	3.40	1.52	-0.88	
16	Overall, I am satisfied with this system.	3.00	1.22	-1.83	

An analysis of results reveals some consistent patterns. First, with the exception of a single item ("The system gave me error messages that clearly told me how to fix the problems"), all of the item means were better than the scale midpoint value of 4.0. Second, the users reported favorable attitudes towards the portal. Specifically, 8 of the 16 knowledge base items reached statistical significance. An analysis of the users' qualitative comments included the desire for greater clarity of terminology.

Based on the usability study, changes were made to the knoweldge base. These changes include some terminology used, such as changing the framework component names (e.g., performance to skill/task, performer to user, and effectiveness to outcome). Additionally, short summary write-ups were created for each of the studies in the knoweldge base to provide users with a quick overview of the study. The study abstracts rarely described all components of the framework for a particular study, so the abstracts were insufficient for a user to understand the relevance to their query based on framework components.

5. Summary and Conclusion

The current state of AR/VR technology effectiveness studies is fragmented and disorganized. Some of the reasons for this disorganization is the inconsistent use of terms (e.g., using the same term to describe different things or different terms to describe the same thing) that may be due to the studies being done by researchers in disparate fields (e.g., technology developers, education/training researchers, and domain specialists and trainers). The objective of this project was to develop a framework for organizing a knowledge base of AR/VR studies in both the current and future literature so users can identify what's important and meaningful.

The framework, based on an initial review of the literature and discussions with subject-matter experts, helps to highlight information that may influence or impact the effectiveness of AR/VR training systems. The framework has four dimensions of an AR/VR training effectiveness study: technology, skill/task, users, and training use. In addition, the outcome of the effectiveness study is the fifth component of the framework. All parts should be described clearly in a training effectiveness study to foster a reader's understanding of the context for the AR/VR implementation; they also may indicate when the results are likely to generalize (or not) to other training contexts.

The content for the knowledge base was developed by having a researcher review a study and summarize it using all components of the framework. A second researcher then reviewed the knowledge base entry and made edits or additions where needed. This knowledge base with organized content allows users to filter for one or more components that address their specific queries.

A. Summary of Findings from the Knowledge Base

To date, the knowledge base includes 64 studies with the expectation for additional studies to be added over time. A range of queries of the current knowledge base reveals insights into the growing corpus of studies that have assessed the effectiveness of AR and VR technology for training, education, and job performance.

The majority of system comparison studies found that AR/VR was more effective than traditional training. Several other studies had mixed results indicating that AR/VR was about the same as traditional training, particularly when new prototype AR/VR systems are compared to well-established training systems. Very few AR/VR systems compared to traditional training were less effective.

There were numerous studies geared to inform the design of AR/VR that did not compare the system to traditional training. They compared components of AR/VR (e.g., with or without a particular system feature) to other feature variants of the same system. These studies, show what may make AR/VR better or worse or when to use particular features. Examples of insight gained from these types of studies include the importance of instructional feedback (i.e., feedback that guides a user to improve performance) and how the content of what is displayed (e.g., mapping directions or highlighting salient cues) can help improve performance.

Each dimension of the framework highlights different characteristics of studies. Both individually and collectively, the dimensions enable knowledge base queries to emphasize a knowledge base user's interests. Some may care more about the technologies, others about the skills/tasks being trained, and so on. Each framework dimension, illustrated in what follows, shows the value that such organization brings to studies.

There are many different types of technologies that comprise AR/VR. The most frequently studied technology in the knowledge base was an occluded HMD; the use of computer monitors and see-through HMDs also were fairly common for displaying visual content. There were also several technologies that tracked the user's actions (e.g., moving their head or hands) or tracking the equipment or objects the user manipulated. However, there was no clear indication of why one type of technology worked and another didn't. Also, the breadth of technologies used shows how AR/VR is not a singular entity, but a class of various technologies that help to provide an extension to reality for users.

In the knowledge base, the two most frequent training domains or skills/tasks that AR/VR addresses are maintenance/assembly and medical, followed by flight and military tasks. With maintenance and medical, there are many examples of AR/VR being more effective than traditional training. With flight training, the most common outcome was mixed results, but those comparisons were usually against well-established flight simulators that were already being used for training pilots. With the studies of military tasks, most of the findings were component evaluations, suggesting that this area is still working to develop how this technology can potentially be used effectively.

The skills/tasks being trained can also be characterized by the generic type of physical (e.g., fast or slow motor actions) or cognitive actions (e.g., decision making, acquiring knowledge) that the user is training to complete. By querying based on a set of generic skill/task descriptors, a knowledge base user could find additional studies that might be relevant to their interests. Another benefit of generic skill/task descriptors is that they may allow knowledge base users to differentiate between studies in broad training domains such as military tasks used to train physically dominated fast combat actions versus cognitive heavy strategic planning.

The training methodology is an essential ingredient to consider when assessing the effectiveness of an overall training system. There are many ways to incorporate AR/VR into training such as a practice environment that may have automated feedback or live-instructor feedback. Using the system as a practice environment was the most common in the knowledge base with several studies including different kinds of feedback to help guide a user to improve performance. Additionally, some systems go beyond training in that they are expected to be used on the job as an aid to a person doing a task.

The user (i.e., person being trained) is also a key component to any effectiveness study. Training a person with some experience in a particular domain is very different than training a complete novice. A large number, just a little less than half, of the studies in the knowledge base did not provide information on the level of experience of the user in the training domain or their prior use of AR/VR technology. The importance of including user experience is highlighted by studies that included both experienced and novice users; those with either prior skill/task experience or AR/VR technology experience performed better than trainees without it.

In general, knowledge base queries allow their user to organize studies by their important features and components. They provide flexibility for the user to choose a search strategy from a wide range of options such as technology types, skill/task types, types of metrics, and user experience levels with task and technology. An essential part of that flexibility is to mix and match search terms, to cut across components, and find their intersection (e.g., technology with task type with outcome). And, an added benefit is to identify research gaps where few or no studies are found. The principal limitation is the number of studies in the knowledge base which need to grow as the user community identifies more that are relevant and interesting. The time needed to code articles and insert them into the knowledge base is not trivial, so there is still considerable work to be done in this area.

B. Conclusion

The science of building AR/VR training systems is still developing. This report creates a foundation for collecting and organizing information from AR/VR studies and enhancing the development of knowledge about training effectiveness. That foundation and information organization provide a consistent terminology and a framework to help users sift through and understand the nuances of studies and their effectiveness for training. The framework components feature what the technology is, who is the user, what are they being trained to do, and how the technology is integrated into a training methodology (use). An initial knowledge base of 64 studies that employs the framework should grow to provide an ever-improving resource for users to understand how AR/VR can be effective.

Next steps for this effort may include inclusion of the knowledge base as part of an AR/VR asset repository (Beaubien et al. 2022). This would enable a broader set of knowledge base users that would have access to the contents and could provide additional feedback for further development. Additionally, the intent is to transition the knowledge base to a crowd-sourced model of content development, rather than a researcher-sourced model of content development.

C. Acknowledgements

This effort was funded by the Air Force Research Laboratory, Airman Systems Directorate, Warfighter Interaction and Readiness Division (AFRL/RHW) in support of the Commander Air Combat Command's (COMACC) migration to Proficiency-Based Training. The authors owe particular thanks to those who contributed to the development of the knowledge base and its contents: Dexter Fletcher, Aida Muzo, Adam Sequera, Matthew Brinkley, Christian Dobbins, Janna Mantua, Brad Boyerinas, Isaac Chappell, Matthew Trowbridge, and Felicia Sallis-Peterson. The authors would also like to thank reviewers John Morrison and Sue Numrich of IDA as well as Jeff Beaubien from Aptima for their insightful comments and suggestions to improve the report.

Appendix A.
Codebook

The tags listed below are structured in accordance with the five primary questions of the framework: a) what is the technology, b) how is the technology being used, c) who is the user, d) what is the performance or training domain, and e) was it effective? At least one tag should be used for each of the five primary questions; however, multiple tags could apply to a given question and all relevant tags for an article should be used. The tags should represent the important and salient features of the instantiation being characterized. This may take some interpretation of what is presented in the system evaluation because the author did not write the effectiveness report based on the framework.

The intent of the tagging process is to help identify relevant effectiveness studies that will inform the knowledge base user for a particular query. This approach allows us to be more systematic in our analysis of the literature and to capture more of the landscape than a simple keyword search (i.e., using CTRL +F + keyword in the spreadsheet). It will also allow us to identify instances with particular combinations of characteristics (tags) that a keyword search would not allow us to do—e.g., wanting to learn about empirical evidence on handheld devices and slow motor movement tasks. With that intent in mind, select the tags that appropriately describe the effectiveness study.

Technology

These tags describe the AR/VR system that is being assessed in the article.

Audio

- tech_audio_headphones – audio information is provided to the user through headphones.
- tech_audio_speakers – audio information is provided to the user through speakers not worn by the user.
- tech_audio_record – audio information is captured; it could be through a microphone, phone, or other type of device. The audio could be recorded or just transmitted to another user or the system.

User Tracking

- tech_user_body_tracking – system tracks the user’s body location and/or orientation (i.e., direction the body is facing, e.g. torso).
- tech_user_eye_tracking – system tracks where the user’s eyes are looking and/or what they are looking at.
- tech_user_head_tracking – system tracks the orientation of the user’s head.
- tech_user_hand_tracking – system tracks the user’s hand location and/or orientation (at least one hand). This system could also determine what the user is doing with their hand/s (e.g., grasping, rotating, gesture). This would include a handheld controller or glove that provides the hand movement information.

Environment Tracking

- tech_environment_geolocation – system is tracking the location of objects (e.g., user or other objects) in the geographical location. This could be on a real (i.e., actual location) or a virtual (created) mapping.
- tech_environment_marker_tracking – system uses tags or markers in the environment to understand the location/orientation of objects and people in the environment.
- tech_equip_tracking – system tracks a specific piece of equipment or tool. From this tracking, the system gains an understanding for what the person is doing to the equipment (i.e., fixing a device, inserting a component).
- tech_environment_object_discriminate – system identifies at least some objects in the environment, and/or is able to assign some characteristics to the object (e.g., vehicle, person other than user, real-world signs).

Haptic Feedback

- tech_haptic_body – system provides some haptic feedback through a body-worn device (e.g., chest or back).
- tech_haptic_controller – system provides haptic feedback through a controller device that the user is holding (e.g., joystick, augmented hand tool, handheld weapon).
- tech_haptic_gloves – system provides haptic feedback through a glove worn by the user.
- tech_haptic_other – system provides haptic feedback through some additional means not described by the other haptic feedback tags.

Software

- tech_sw_equipmech_model – system includes a dynamic model of a piece of equipment or mechanical device that a user may be working on. For medical training, a model of the patient may be considered an 'equipmech' model. The system should use the model to represent how things change (e.g., replace part, insert needed in the correct location, tighten screw).
- tech_sw_instruct_model – model includes the appropriate behavior the user should be doing and provides some level of instruction when the user performs correctly or incorrectly.
- tech_sw_intel_avatars – there are virtual characters in the environment that the user can interact with at some level. The level of interaction can be relatively simplistic (e.g., if the user shoots the avatar the avatar falls down) or it could be quite sophisticated in that the avatar might engage in two-way conversation.
- tech_sw_rendering – system builds (renders) a virtual environment that is updated in real-time or nearly real-time (i.e., milliseconds of lag time).

Visual Display

- Computer display
  - tech_visual_computer – visual information is provided to the user through a computer monitor or some type of TV screen that is relatively fixed. It could be a computer monitor that is on a desk or in some type of simulator set-up. It could also be fixed to a wall or attached to some other surface.
  - tech_visual_handheld – visual information is provided on a screen that is held by the user and the user can move the screen. This could be a relatively small screen as with a phone or a larger screen with a handheld tablet device.
- Head-Mounted Display (HMD)
  - tech_visual_HMD_occluded – visual information is provided to the user through a system that is attached to the user's head so it moves in unison with the user and all of the visual information is provided on the screen (i.e., the user cannot see the real world, except via the display system).
  - tech_visual_HMD_see_through – visual information is provided to the user through a system that is attached to the user's head so it moves in unison with the user and some of the visual information is provided on the display but the user can simultaneously also see the real world.
- tech_visual_projection – visual information is projected onto a surface so that the user can see the projected image as if it was part of the object/environment being viewed.

Technology Not Discussed

- tech_NA – there is no mention or description of the technology (e.g. if the article is qualitative, or a rationale).

Skill/Task

These tags are to describe the skill, task, or job that the user is supposed to learn or complete with AR/VR system. These indicate the 'performance' expected by the user that may be aided by use of the technology.

Auditory

- perf_auditory_comprehend_speech – the task requires the user to comprehend spoken words (e.g., hearing verbal directions, having a conversation). The speech that the user has to comprehend does not have to be produced by the system.
- perf_auditory_detection – the task requires the user to detect a sound (e.g., hearing a potential mechanical problem based on sound, hearing an alarm or system warning).
- perf_auditory_orient – the task requires the user to turn towards the sound or identify the location/direction of what may be making the sound.

Cognitive

- perf_cog_acquiring_knowledge – the task requires the user to acquire knowledge by gaining and/or understanding and/or learning new facts. Demonstration of knowledge isn't required.
- perf_cog_decision_making – the task requires the user to make one or more decisions over the course of the task. This could include comparing at least two options and selecting one; problem solving; planning actions or determining a sequence of actions to achieve a goal; or calculating a solution. This could also include situational awareness, where a person has to both understand the current state and a potential future state.
- perf_cog_immersion – the user has the perception that they are actually engaging in and/or truly experiencing what is simulated. This tag could also be used for studies that are comparing presence/immersion in systems while conducting a task.

Sensorimotor Coordination

- perf_sensori_eye_motor_coordination – the task requires the user to perform actions that combine vision and moving one's body/limbs (e.g., catching a thrown ball, moving flight controls based on visual sense to fly an aircraft).

- perf_sensori_touch_motor_coordination – the task requires the user to perform actions that are guided by a user's sense of touch (e.g., palpate patient's body for medical exam).

Motor Skills

- perf_motor_speed_fast – the task requires the user to perform an action with quick movements (e.g., running, quickly orienting to immediate threat as during combat, juggling, fast typing, rapidly assembling a rifle).
- perf_motor_speed_slow – the task requires the user to perform an action with slow movements (e.g., walking, putting a puzzle together, turning a wrench).
- perf_motor_manipulation_fine – the task requires the user to perform an action that requires fine motor skills (e.g., fixing something with tweezers, playing violin, turning a dial with fingers, needle insertion).
- perf_motor_manipulation_gross – the task requires the user to perform an action with gross motor movements (e.g., lifting a large object, walking/running, moving feet to orient body towards oncoming threat).

Visual

- perf_visual_discriminate– a primary part of the task is for the user/performer to detect or identify something visually (e.g., light on/off, friend/foe designation, forward observer detecting a target).
- perf_visual_reading – the task requires the user to visually read and comprehend text.
- perf_visual_tracking_move – the task requires the user to visually track a moving object.

Training or Performance Not Discussed

- perf_NA – There is no discussion or mention in any detail about the task or performance.

Training-Technology Integration

These tags describe how the AR/VR system is used to train or perform a task/job, or is integrated into a training course/curriculum.

Extra Information (e.g., vision enhancement)

- use_extra_symbols – system provides labels or symbols to objects that provide additional information than would be available in the real-world environment (e.g., labeling an object with the name and some additional characteristics of the object, highlighting objects, arrow pointers).

Job Aid

- use_job_aid – system provides additional information to help a person complete a job (e.g., vehicle) or skill (e.g., aim a weapon, land a plane, or replace a part). The system could be used in a real-world scenario to accomplish a job skill (e.g., a system that could be brought into an operating room and used to perform a specified procedure). Does not include if performance is conducted on plastic models.

Technology Evaluation

- use_tech_eval_design – the AR/VR technology is used to help design something (e.g., computer-aided design).
- use_tech_eval_general – the AR/VR technology is used to test a model of a yet to be built system/device to see how well it might work. For example, a new car is being designed and a virtual version of that car is developed and a person uses AR/VR technology to test drive the car.

Training

- use_training_automated_feedback – system provides instructional feedback to the user to help them learn a task/skill. This could include a description of how to perform the task, then having the user/trainee perform the task, and the system provides an indication of what they did right/wrong or how they could improve their performance.
- use_training_instructor_feedback – system provides an environment where a user can learn/practice task and the real-person instructor provides feedback. The system does not provide feedback for improvement.
- use_training_practice – system provides an environment where a user can practice a task/skill. This tag would be appropriate if there is a real-person instructor that is teaching the task/skill and the AR/VR system just enables the user/trainee to practice the skill.

Technology Use Not Discussed

- use_NA – there is no mention of how the technology is used.

Technology Users

Technology Users – these tags describe the studied user of the AR/VR system. It is divided by how skilled the user is in the performance task and with similar AR/VR systems.

Skill Performance Level

- user_task_novice – the user is a relative beginner in the skill being trained or performed with the assistance of the AR/VR system.
- user_task_experience – the user has some experience in the skill being trained or performed with the assistance of the AR/VR system. Experience level can range from some experience to expert level.

Technology Performance Level

- user_tech_novice – the user does not have experience with AR/VR technology.
- user_tech_experience – the user may have some experience with AR/VR technology, either the system being tested or systems that share some characteristics with the one being tested. Experience level can range from some experience to expert level.

Skill Performance Level Not Discussed

- user_task_NA – there is no mention of the user's prior experience performing the task described.

Technology Users Not Discussed

- user_tech_NA – there is no mention of the user's prior experience using the technology described.

Effectiveness

These tags are to describe the method used to assess if the system was effective or not. This focuses on the type of data that were collected and used to indicate if a particular AR/VR instantiation was effective or not.

Empirical – Quantifiable Measure

- empirical_measure_perform – there is a numerical value placed on the physical performance of the user and the value is objectively determined. Examples include system-generated scores, percent actions correct, and number of errors.
- empirical_measure_cognitive - there is a numerical value placed on a cognitive outcome by the user and the value is objectively determined. Examples include percent questions answered correctly, number of correct planning steps completed, and ratio of correct/incorrect facts recalled.
- empirical_rating_performance – the performance of the user is rated on a subjective scale. For example, using a 10-point Likert scale the instructor may indicate how good/bad a user's performance was.

- empirical_rating_useability – the users or observers rate the usability of the system to accomplish an objective. A standard usability scale is the System Usability Scale (SUS), but there are others, that addresses if the system can be used to achieve objectives, how much effort is required to achieve objectives, and was the user experience satisfactory or not.
- empirical_rating_workload – subjective ratings of system use to perform a task. A common workload scale is the NASA TLX that describes the level of mental, physical, and temporal demand exerted to achieve a level of performance, along with a rating of potential frustration during the task.

Opinion – Narrative Description

- opinion_expert – an expert in the field or the instructor provides a verbal/text description of how well the system worked in particular circumstances. An example could be statements like, “The user’s benefited by practicing with the AR system.” The level of expertise should be above a standard user that may be moderately experienced in the domain.
- opinion_non-expert – a verbal/text description by a person without extensive experience in the domain, describing how well a system worked as intended to accomplish a task. Most common non-experts providing their opinion would be users of the system while testing it.

Appendix B.
Analyzing Effectiveness Studies

Purpose

A key objective of the knowledge base is to organize the knowledge and make it easy to search. Initial reviews of the AR/VR research literature found inconsistent use of terms across studies, incomplete information on the context for their use, potential for over-generalizations of findings, and an assortment of ways to measure effectiveness. To organize the information, we developed the framework shown below (Figure B-1) and used it to develop a compatible and searchable knowledge base.

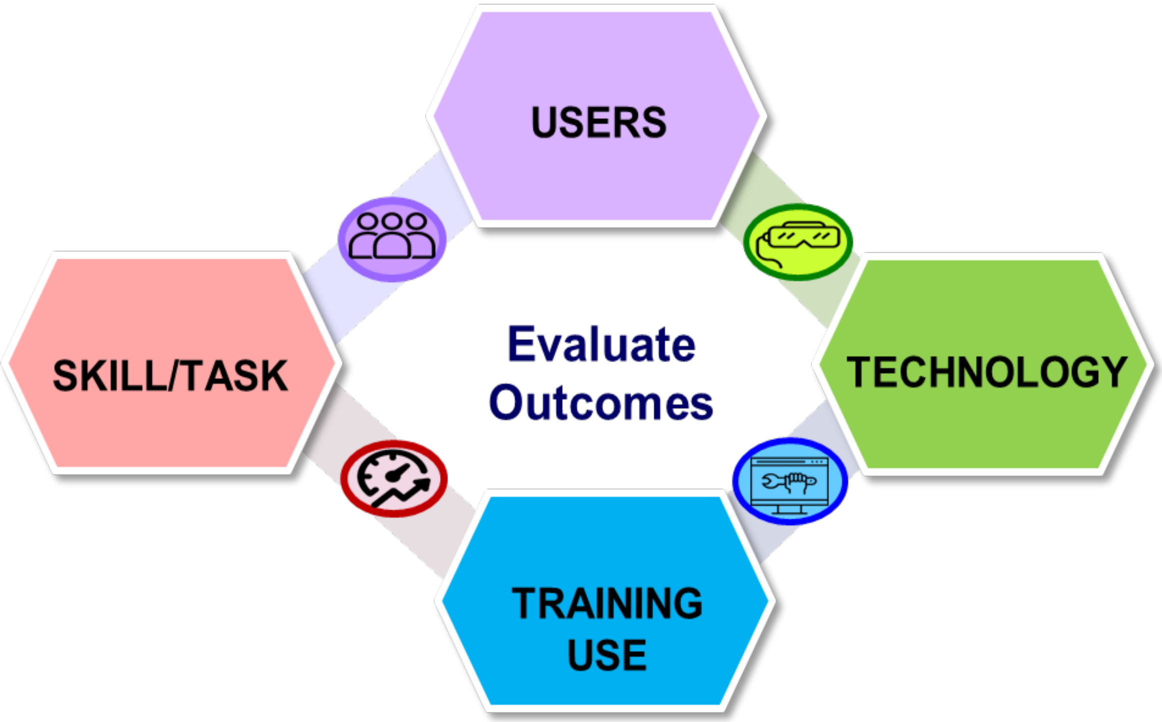


Figure B-1. Graphical depiction of the framework.

Introduction to Framework and Knowledge Base

To employ the framework, there are five questions to answer for each study included in the knowledge base:

1. What is the technology being used?

2. What is the skill/task being trained or performed?
3. How is the technology used for training or integrated into a situation/course?
4. Who is the user?
5. What is the outcome of AR/VR use (i.e., evaluation of effectiveness)?

Each of those questions is answered using “specific” and “general” descriptions, with summaries of this information inserted into a template (Figure B-2), reviewed, and then copied into the knowledge base. The “specific” summarizes the instantiation of the answer closely adhering to what in particular was done, how it was done, and its outcome. The “general” is a description of the characteristics and utility of the answer. For example, a specific technology could be the Microsoft HoloLens while the general description may include the device’s characteristics (e.g., see-through HMD with head-tracking sensors, a camera pointed in the direction of the user’s gaze, headphones, and a microphone), and how they might be employed in the application. The expectation is that reading specific information in an answer will provide essential facts and that general information will highlight their characteristics and utility. There may, of course, be some overlap in the specific and general descriptions. The ultimate goal is to allow users of the knowledge base to extrapolate information from a single study to their own unique use case. Users can explore the nuances of AR/VR literature and move beyond only answering: What was done and is it effective? Specific and general descriptions help to specify the multi-faceted characteristics of AR and VR studies.

Citation/Title	First author’s last name Year of study Title of Study Huang et al. 2018. The use of augmented reality glasses in central line simulation: "See one, simulate many, do one competently, and teach everyone."
Reference	Authors and initials (Year of study) Title of Study Citation Huang, C. Y., Thomas, J. B., Alismail, A., Cohen, A., Almutairi, W., Daher, N. S., Terry, M. H., & Tan, L. D. (2018). The use of augmented reality glasses in central line simulation: "See one, simulate many, do one competently, and teach everyone." Advances in Medical Education and Practice, 9, 357–363.
Abstract	Copy from study
Tech Specific	The study used # technology systems: [list systems if included in the article, and in this order: visual, audio, tracking, haptic]. For each system include only features that were used: [Name of Technology system 1; features of Technology system 1]; [Name of Technology system 2; features of technology system 2], … [and more systems if relevant]. The software used was: [name software] and the study used [list software features used]. Notes: Include the features of the system that are used in addition to just naming the device (e.g., HoloLens with a see-through visual display, forward-facing camera, microphone array). Just need to describe the features used in the system assessed, no need to include features that are not used in this instantiation.
Tech General	

Tech Tags	
Performance Specific	The task(s) in this study required the user to [list main task(s)]. This included the users performing [list sub tasks]. [Describe any additional task-specific details that are important for the study.] Notes: This could include the name of the overall job/task as well as specific steps in the task. For example, it could be something like – “fly an F-16, which includes taking off, maintaining level flight, turning, and landing the plane.” If there is a broad category that may provide context for the performance (e.g., medical, maintenance, military) that can be included.
Performance General	
Performance Tags	
Training Use Specific	The technology used in this study helped the user do [list task]. The training/instruction included [list instruction, e.g., directions, cues for action, extra text displayed, performance feedback]. Notes: Describe how the technology is used for training or performance aiding in this study. If part of a course, provide a brief description (i.e., type or name of course such as beginning pilot lessons; theme of course such as administering shots during first-year of nursing school; purpose of course such as job aid for fixing a tank’s tread). In short, answer what information the technology system provides to the user, both the virtual entities and the instructional information. How does the instructor incorporate the technology into teaching or performance aiding, or is there no instructor?
Training Use General	
Training Use Tags	
Performer Specific	This study had # participants (# male, # female). The users had [novice/expert] experience with the task and [novice/expert] experience with the technology. [List any additional salient features of the participants for the study, e.g., if familiarity with AR/VR systems was assessed, study specific age range, who the system might be used by.] Notes: NA
Performer General	
Performer Tags	
Effective Specific	Effectiveness summary: [describe bottom line take away]. The empirical measures recorded in this study were [list measures, e.g., accuracy]. Subjective feedback [was/was not] recorded. For the empirical measures: [describe results with statistics]. For the subjective feedback, participants [describe feedback]. Notes: If the AR/VR system was compared to another method of training, describe the comparison. Describe any statistical test results, the magnitude of

	any differences (e.g., effect size, probabilities), and explain any problematic aspects of the results. Indicate if there was a transfer of training demonstration (i.e., Did the performance in the training context improve performance in the real-world context of the training objective?) Describe ratings evidence.
Effective General	
Effective Tags	
Notes	

Figure B-2. Template for analyzing and summarizing studies.

The primary sections in the review template are for Technology (Question 1) and Skill/Task (Question 2) because these may be the most salient for likely queries. What is the Training/Performance Use (Question 3) and Outcome (Question 5) are the next two most important sections, followed by Users (Question 4) as least critical but good to know.[9] An "NA" (i.e., not applicable) option (available for all five sections) is used only as a last resort when a study has no relevant information.

In addition to the narrative descriptions, each section of the template is tagged with descriptors based on the five questions that accompany the framework. The tags are codes for important and salient features to use in searching the knowledge base. The tags help identify studies that share common characteristics. The minimum, although there is no strict number, is one tag per component of the framework, but enough should be applied to represent the study, and the most salient tags should be the first ones inserted.

Steps in Analyzing a Study for the Knowledge Base

The analysis of a study for the knowledge base first requires determining its relevance and then, for those that qualify, filling in the template. The process may also require some additional information gathering to supplement what the study reports such as technology characteristics.

1. A quick review of the abstract or full paper determines if it qualifies for the knowledge base (i.e., an AR or VR effectiveness study with evaluation criteria or evidence to support some outcome). If "yes," continue entering it into the review template following steps 2 through 9 below; if "no," then use a different study if you have one. Exclusion criteria include studies where technology users are classified as children (i.e., younger than 18).

[9] The sample used in the study may not be representative of the expected population for use. For example, the study may involve college students but the ultimate use for the system may be professionals in a particular occupation.

2. Use the template to record standard study information (Citation/Title, Reference, and Abstract). This information can be extracted (often using copy and paste) from the study. The citation/title information is part of the filename for the knowledge base review using the format: first author's last name followed by date of publication and the study's title (e.g., "Belanich 2016 Assessment of training effectiveness with augmented reality and augmented virtuality").

3. Determine the technology being used and fill in the appropriate cells on the template.

    a. Specific – could include the make and model of the device, a description of the prototype if it is not commercial off-the-shelf (COTS). Include the features of the system that are used in addition to just the name of the device. For example, HoloLens has a see-through visual display, inertial sensors, forward-facing camera, microphone array, and more, but all of those features may not be used in a particular study. NOTE: Only describe the features used in the system assessed.

    b. General – describe in narrative the characteristics of the device as you would if explaining what "it" is or the capabilities "it" has to a person who might be interested in an application similar to the one studied. What general features does the device have as used in the study? It may be helpful to think of what tags from the codebook are relevant for the system, and then use those features in this general description.

    c. Enter Technology tags for the study.

4. Determine the task/skill being trained or performed with the technology (i.e., performance to be achieved), and fill in the appropriate information for the knowledge base. What is the trainee/user doing?

    a. Specific – describe the learning objective. This could include the name of the overall job/task as well as specific steps in the task. For example, it could be something like – "fly an F-16, which includes taking off, maintaining level flight, turning, and landing the plane." A broad category that gives context or a domain of performance (e.g., medical, maintenance, military) may be useful to include.

    b. General – describe in narrative the salient characteristics of the performance using general terms from visual, auditory, cognitive, and psychomotor types of descriptors, or other general descriptive terms that would allow a reader to understand the performance and how it may be similar to other types of activities. Using the example of "fly an F-16" above, the general description might be "cognitively determine the appropriate steps in a pre-flight checklist," "physically manipulate the flight controls based on visual determination of how the plane is responding (i.e., this would require hand-eye psychomotor coordination)."

    c. Enter Performance tags for the study.

5. Determine how the technology is being used or how it helps a person learn the task (i.e., how it is being incorporated into a course or curriculum, what training content it provides a learner, or how it is being tried out as a job aid).

   a. Specific – describe how the technology is used for training or performance aiding in this study. If part of a course, provide a brief description (i.e., type or name of course such as beginning 'pilot lessons,' theme such as 'administering shots during first-year of nursing school,' purpose such as 'learn how to fix a tank's tread'). Then, include how the system provides instructional information (e.g., directions, system-generated feedback, practice environment) to the user. A distinction can be made between instructional information provided by the system and by an instructor. Is a real live person (instructor) needed as part of the total instantiation (i.e., does the instructor incorporate the technology into teaching or performance aiding or is that part of the technology)?

   b. General – describe in brief statements the salient characteristics of the way the system is being used for training or performance. For example, does the system provide automated feedback or does it require an instructor to provide feedback to the trainee, is it used for practice only, or is it a job aid to help a person complete the task?

   c. Apply the Technology Use tags for the study.

6. Determine the user characteristics (i.e., a novice or an expert in the technology and the task/skill). Are there particular characteristics that make the study users unique or uncommon?

   a. Specific – describe the users. Are they students of a particular education level or occupation? Have the students already built up a foundation of relevant knowledge before using the system (e.g., third-year nursing students, pilots with at least 1,000 flying hours)? Also, how much experience do they have in using similar AR/VR technology?

   b. General – describe in brief statements the salient characteristics of the users. Are they domain novices or experts?

   c. Apply the User tags for the study.

7. Determine the study's outcome (i.e., how well did the system work and what is the evidence for its success?), and fill in the appropriate parts of the template.

   a. Specific – describe the evidence for the study's outcome beginning with an overall synopsis and progressing to more details. If the AR/VR system was compared to another method of training, describe the comparison. Describe any statistical test results, the magnitude of any differences (e.g., effect size), and explain any problematic aspects of the results. Indicate if there was a transfer of training demonstration (i.e., did the performance in the training context improve performance in a real-world context?).

   b. General – describe how the evidence may indicate that this study could contribute to generalizable principles. For example, a handheld tablet improves the time a novice mechanic takes to learn the (psychomotor) skills necessary for maintaining a new vehicle (given prior knowledge of the vehicle).

   c. Apply the Outcome/Effectiveness tags for the study.

8. Develop a summary statement.

   a. Summarize study in approximately 100 words or less for the knowledge base user to quickly capture the key points of the study. The summary includes the five components (technology, skill/task, training use, user, and outcome) to provide a short synopsis of what was done. The order or level of detail from the components may vary based on their importance within the study.

   b. Overall, when writing a summary, think about the reader who wants a quick overview with key information. Avoid jargon as much as possible so that the summary is easily readable. For example, *"An Epson BT 200 binocular see-through heads-up display (i.e., smart glasses) was used to provide extra visual data of flight-related information (gauges) to licensed pilots as they executed a number of challenging flight-related scenarios in a simulated practice environment. When compared to using a standard heads-down display, the smart glasses users had fewer errors (i.e., missed signals and path deviation errors) than the head down display condition. There were no differences for subjective workload and reaction time."*

9. There also are places for "Notes" in the template for any additional information that clarifies narrative entries.

10. Read through the information to check your work. Ask someone else not familiar with the study to read your template narratives and tags for clarity. Revise as needed.

    a. Could a person reading the information (specific and general) understand the study without consulting the publication?

    b. Will the tags for each of the five questions appropriately identify this article so a query of the knowledge base should find it?

11. Finally, copy the template into the knowledge base.

References

Note – the references with an asterisk are currently in the knowledge base.

* Abidi, M. H., A. Al-Ahmari, A. Ahmad, W. Ameen, and H. Alkhalefah. 2019. "Assessment of Virtual Reality-based Manufacturing Assembly Training System." *International Journal of Advanced Manufacturing Technology* 105 (9): 3743–3759.

* Aebersold, M., T. Voepel-Lewis, L. Cherara, M. Weber, C. Khouri, R. Levine, and A. R. Tait. 2018. "Interactive Anatomy – Augmented Virtual Simulation Training." *Clinical Simulation in Nursing* 15: 34–41.

* AlNajdi, S. M., M. Q. Alrashidi, and K. S. Almohamadi. 2018. "The Effectiveness of Using Augmented Reality (AR) on Assembling and Exploring Educational Mobile Robot in Pedagogical Virtual Machine (PVM)." *Interactive Learning Environments* 26: 1–17.

Azuma, R. T. 1997. "A Survey of Augmented Reality." *Presence: Teleoperators & Virtual Environments,* 6(4), 355–385.

Azuma, R., Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. 2001. "Recent Advances in Augmented Reality." *IEEE Computer Graphics and Applications*, 21(6), 34–47.

* Bach, B., R. Sicat, J. Beyer, M. Cordeil, and H. Pfister. 2017. "The Hologram in my Hand: How Effective is Interactive Exploration of 3D Visualizations in Immersive Tangible Augmented Reality?" *IEEE Transactions on Visualization and Computer Graphics* 24(1): 457–467. https://doi.org/10.1109/TVCG.2017.2745941.

* Bailey, S. K., C. I. Johnson, B. L. Schroeder, and M. D. Marraffino. 2017. "Using Virtual Reality for Training Maintenance Procedures." In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*.

* Bailey, S. K. T., and C. I. Johnson. 2020. "A Human-Centered Approach to Design in Gestures for Natural User Interfaces." In: Kurosu, M. (eds) *Human-Computer Interaction. Multimodal and Natural Interaction. HCII 2020. Lecture Notes in Computer Science()* vol 12182. https://doi.org/10.1007/978-3-030-49062-1_1.

* Balian, S., S. K. McGovern, B. S. Abella, A. L. Blewer, and M. Leary. 2019. "Feasibility of an Augmented Reality Cardiopulmonary Resuscitation Training System for Health Care Providers." *Heliyon* 5(8), e02205.

* Batmaz, A. U., X. Sun, D. Taskiran, and W. Stuerzlinger. 2020. "Eye-Hand Coordination Training for Sports with Mid-air VR." *Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST*.

* Batmaz, A. U., and W. Stuerzlinger. 2021. "The Effect of Pitch in Auditory Error Feedback for Fitts' Tasks in Virtual Reality Training Systems." In *2021 IEEE Virtual Reality and 3D User Interfaces (VR*), 85–94. https://doi.org/10.1109/VR50410.2021.00029.

* Batdi, V., and T. Talan. 2019. "Augmented Reality Applications: A Meta-analysis and Thematic Analysis." *Turkish Journal of Education* 8(4): 276-297.

Beaubien, J. M., W. Bennett Jr., R . B. Ayers, R. Keithley, K. Audrain, and J. Belanich. 2022. "Development of a Searchable, Web-Based Repository for Sharing ARVR Training Assets." I/ITSEC Paper No. 22252. In *Proceedings of the 2022 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Arlington, VA: National Training and Simulation Association.

* Bhagat, K. K., W. K. Liou, and C. Y. Chang. 2016. "A Cost-Effective Interactive 3D Virtual Reality System Applied to Military Live Firing Training." *Virtual Reality* 20(2): 127–140.

* Bork, F., A. Lehner, U. Eck, N. Navab, J. Waschke, and D. Kugelmann. 2021. "The Effectiveness of Collaborative Augmented Reality in Gross Anatomy Teaching: A Quantitative and Qualitative Pilot Study." *Anatomical Sciences Education* 14(5): 590–604.

* Brizzi, F., L. Peppoloni, A. Graziano, E. Di Stefano, C. A. Avizzano, et al. 2018. "Effects of Augmented Reality on the Performance of Teleoperated Industrial Assembly Tasks in a Robotic Embodiment." *IEEE Transactions on Human-Machine Systems* 48(2): 197–206. https://doi.org/10.1109/THMS.2017.2782490.

* Chalhoub, J., and S. K. Ayer. 2019. "Exploring the Performance of an Augmented Reality Application for Construction Layout Tasks." *Multimedia Tools and Applications* 78: 1–24. https://doi.org/10.1007/s11042-019-08063-5.

* Clifford, R., T. McKenzie, S. Lukosch, and R. Lindeman. 2020. "The Effects of Multi-sensory Aerial Firefighting Training in Virtual Reality on Situational Awareness, Workload, and Presence." *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*: 93–100. https://doi.org/10.1109/VRW50115.2020.00023.

* Cooper, N., F. Milella, C. Pinto, I. Cant, M. White, and G. Meyer. 2018. "The Effects of Substitute Multisensory Feedback on Task Performance and the Sense of Presence in a Virtual Reality Environment." *PLoS ONE* 13(2). https://doi.org/10.1371/journal.pone.0191846.

* Corelli, F., E. Battegazzorre, F. Strada, A. Bottino, and G. P. Cimellaro. 2020. "Assessing the Usability of Different Virtual Reality Systems for Firefighter Training." In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2: 146–153.

* Dalladaku, Y., J. Kelley, B. Lacey, J. Mitchiner, B. Welsh, and M. Beigh. 2020. "Assessing the Effectiveness of Virtual Reality in the Training of Army Aviators." In *Proceedings of the 2020 Annual General Donald R. Keith Memorial Capstone Conference*, 40. New York, NY.

* Deshpande, A., and I. Kim. 2018. "The Effects of Augmented Reality on Improving Spatial Problem Solving for Object Assembly." *Advanced Engineering Informatics* 38: 760–775.

* Dhalmahapatra, K., J. Maiti, and O. B. Krishna. 2021. "Assessment of Virtual Reality Based Safety Training Simulator for Electric Overhead Crane Operations." *Safety Science*, 139, art. no. 105241.

* Fan, Y.-C., and C.-Y. Wen. 2019. "A Virtual Reality Soldier Simulator with Body Area Networks for Team Training." *Sensors* (Switzerland), 19 (3), art. no. 451.

Faulkner, L. 2003. "Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing." *Behavior Research Methods, Instruments, & Computers* 35, 379–383 (2003). https://doi.org/10.3758/BF03195514.

Fletcher, J. D., J. Belanich, F. Moses, A. Fehr, and J. Moss. 2017. "Effectiveness of Augmented Reality & Augmented Virtuality." In *MODSIM (modeling & simulation of systems and applications) World Conference*. https://modsimworld.org/papers/2017/ Effectiveness_of_AR_and_VR.pdf.

* Frederiksen, J. G., S. M. D. Sørensen, L. Konge, M. B. S. Svendsen, M. Nobel-Jørgensen, F. Bjerrum, and S. A. W. Andersen. 2020. "Cognitive Load and Performance in Immersive Virtual Reality Versus Conventional Virtual Reality Simulation Training of Laparoscopic Surgery: A Randomized Trial." *Surgical Endoscopy* 34(3): 1244–1252.

* Frutos-Pascual, M., J. M. Harrison, C. Creed, and I. Williams. 2019. "Evaluation of Ultrasound Haptics as a Supplementary Feedback Cue for Grasping in Virtual Environments." In *2019 International Conference on Multimodal Interaction*, 310–318.

* Fussell, S. G., and M. P. Hight. 2021. "Usability Testing of a VR Flight Training Program." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* vol. 65, no. 1: 1124–1128. Los Angeles, CA: SAGE Publications.

Garzón, J., S. Baldiris, J. Gutiérrez, and J. Pavón. 2020. "How Do Pedagogical Approaches Affect the Impact of Augmented Reality on Education? A Meta-analysis and Research Synthesis." *Educational Research Review* 31, 100334.

Garzón, J., and J. Acevedo. 2019. "Meta-analysis of the Impact of Augmented Reality on Students' Learning Gains." *Educational Research Review* 27: 244–260.

* Gierwiało, R., M. Witkowski, M. Kosieradzki, W. Lisik, Ł. Groszkowski, and R. Sitnik. 2019. "Medical Augmented-Reality Visualizer for Surgical Training and Education in Medicine." *Applied Sciences* 9(13).

* Haiduk, P. M. 2017. "A Flight Guidance Display Format on Smart Glasses for Private Pilots." *Aviation Psychology and Applied Human Factors* 7, 2: 66–77. https://doi.org/10.1027/2192-0923/a000119.

* Hale, K. S., G. Campbell, J. Riley, M. Boyce, and C. Amburn. 2019. "Augmented Reality Sandtable (ARES) Impacts on Learning." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* vol. 63, no. 1: 2149–2153. Los Angeles, CA: SAGE Publications.

Haque, S., and S. Srinivasan. 2006. "A Meta-Analysis of the Training Effectiveness of Virtual Reality Surgical Simulators." *IEEE Transactions on Information Technology in Biomedicine*, 10(1), 51–58.

* Hecht, R., M. Li, Quirina M. B. de Ruiter, W. F. Pritchard, X. Li, Venkatesh Krishnasamy, Wael Saad, J. W. Karanian, and B. J. Wood. 2020. "Smartphone Augmented Reality CT-Based Platform for Needle Insertion Guidance: A Phantom Study." *Cardiovascular and Interventional Radiology* 43(5): 756–764.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-analysis*. New York: Academic Press, Inc.

* Hight, M. P., S. G. Fussell, M. A. Kurkchubasche, and I. J. Hummell. 2022. “Effectiveness of Virtual Reality Simulations for Civilian, Ab Initio Pilot Training.” *Journal of Aviation/Aerospace Education & Research* 31(1). https://doi.org/10.15394/jaaer.2022.1903.

* Horner, C., C. K. Padron, and T. Westbrook. 2020. “Evaluating the Use of Augmented Reality for Aircraft Maintenance Training.” *I/ITSEC Proceedings*, Orlando, FL.

* Hou, L., X. Wang, and M. Truijens. 2015. “Using Augmented Reality to Facilitate Piping Assembly: An Experiment-based Evaluation.” *Journal of Computing in Civil Engineering* 29(1), 05014007.

* Huang, C. Y., J. B. Thomas, A. Alismail, A. Cohen, W. Almutairi, N. S. Daher, M. H. Terry, and L. D. Tan. 2018. “The Use of Augmented Reality Glasses in Central Line Simulation: ‘See one, simulate many, do one competently, and teach everyone.’ ” *Advances in Medical Education and Practice* 9: 357–363.

Kaplan, A. D., J. Cruit, M. Endsley, S. M. Beers, B. D. Sawyer, and P. A. Hancock. 2021. “The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-analysis.” *Human Factors* 63(4): 706–726.

* Kwegyir-Afful, E., and J. Kantola. 2021. “Simulation-Based Safety Training for Plant Maintenance in Virtual Reality.” *Advances in Intelligent Systems and Computing*, 1206 AISC: 167–173.

* Kwiatek, C., M. Sharif, S. Li, C. Haas, and S. Walbridge. 2019. “Impact of Augmented Reality and Spatial Cognition on Assembly in Construction.” *Automation in Construction* 108, 102935.

* Leary, M., S. K. McGovern, S. Balian, B. S. Abella, and A. L. Blewer. 2020. “A Pilot Study of CPR Quality Comparing an Augmented Reality Application vs. a Standard Audio-Visual Feedback Manikin.” *Frontiers in Digital Health* 2, 1. https://doi.org/10.3389/fdgth.2020.00001.

* Lerner, D., S. Mohr, J. Schild, M. Göring, and T. Luiz. 2020. “An Immersive Multi-user Virtual Reality for Emergency Simulation Training: Usability Study.” *JMIR Serious Games* 8(3), e18822.

Lewis, James R. 1995. “IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use.” *International Journal of Human-Computer Interaction* 7(1): 57–78.

* Li, F., Y. Tai, Q. Li, J. Peng, X. Huang, Z. Chen, and J. Shi. 2019. “Real-Time Needle Force Modeling for VR-Based Renal Biopsy Training with Respiratory Motion Using Direct Clinical Data.” Agnès Drochon (ed). *Applied Bionics and Biomechanics* 2019 June 25, 14. https://doi.org/10.1155/2019/9756842.

* Liang, C. Y., R. Lascink, and D. H. Klyde. 2020. “Operational Evaluation of a Virtual Reality Parachute Simulator.” In *AIAA Scitech 2020 Forum,* 0669. https://doi.org/10.2514/6.2020-0669.

* Liaw, S. Y., S. W. Ooi, K. D. B. Rusli, T. C. Lau, W. W. San Tam, and W. L. Chua. 2020. "Nurse-Physician Communication Team Training in Virtual Reality versus Live Simulations: Randomized Controlled Trial on Team Communication and Teamwork Attitudes." *Journal of Medical Internet Research* 22(4), e17279.

Limbu, B. H., H. Jarodzka, R. Klemke, and M. Specht. 2018. "Using Sensors and Augmented Reality to Train Apprentices Using Recorded Expert Performance: A Systematic Literature Review." *Educational Research Review*, 25, 1–22.

* Makransky, G., S. Borre-Gude, and R. E. Mayer. 2019. "Motivational and Cognitive Benefits of Training in Immersive Virtual Reality Based on Multiple Assessments." *Journal of Computer Assisted Learning* 35(6): 691–707.

* Marraffino, M. D., C. I. Johnson, D. E. Whitmer, N. B. Steinhauser, and A. Clement. 2019. "Advise When Ready for Game Plan: Adaptive Training for JTACs." In *Proceedings of the Interservice/Industry, Training, Simulation, and Education Conference*.

* McHenry, N., T. Hunt, W. Young, A. Gardner, U. Bhagavatula, B. Bontz, J. Chiu, G. Chamitoff, and A. Diaz-Artiles. 2020. "Evaluation of Pre-Flight and On Orbit Training Methods Utilizing Virtual Reality." In *AIAA Scitech 2020 Forum.* https://doi.org/10.2514/6.2020-0168.

* Mendes, H. C. M., Cátia Isabel Andrade Botelho Costa, N. A. da Silva, F. P. Leite, A. Esteves, and D. S. Lopes. 2020. "PIÑATA: Pinpoint Insertion of Intravenous Aeedles via Augmented Reality Training Assistance." *Computerized Medical Imaging and Graphics* 82, 101731.

Milgram, P., and F. Kishino. 1994. "A Taxonomy of Mixed Reality Visual Displays." *IEICE Transactions on Information and Systems* 77(12): 1321–1329.

* Monteiro P., M. Melo, A. Valente, J. Vasconselos-Raposo, and M. Bessa. 2020. "Delivering Critical Stimuli for Decision Making in VR Training: Evaluation Study of a Firefighter Training Scenario." In *IEEE Transactions on Human-Machine Systems*, 51(2), (2018): 65–74. https://doi.org/10.1109/THMS.2020.3030746.

Moro, C., J. Birt, Z. Stromberga, C. Phelps, J. Clark, P. Glasziou, and A. M. Scott. 2021. "Virtual and Augmented Reality Enhancements to Medical and Science Student Physiology and Anatomy Test Performance: A Systematic Review and Meta-Analysis." *Anatomical Sciences Education*, 14(3), 368–376.

* Muangpoon, T., R. Haghighi Osgouei, D. Escobar-Castillejos, C. Kontovounisios, and F. Bello. 2020. "Augmented Reality System for Digital Rectal Examination Training and Assessment: System Validation." *Journal of Medical Internet Research* 22(8): 1.

* Oberhauser, M., D. Dreyer, R. Braunstingl, and I. Koglbauer. 2018. "What's Real about Virtual Reality Flight Simulation? Comparing the Fidelity of a Virtual Reality with a Conventional Flight Simulation Environment." *Aviation Psychology and Applied Human Factors* 8 (1): 22–34. https://doi/org/10.1027/2192-0923/a000134.

Oprihory, J. 2020. "Pilot Training Next integrated in Experimental Curriculum." *Air Force Magazine*, 103(4), 24–25.

Orvis, K. A., K. L. Orvis, J. Belanich, and L. N. Mullin. 2005. *The Influence of Trainee Gaming Experience and Computer Self-Efficacy on Learner Outcomes of Videogame-based Learning Environments*. Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

* Pettijohn, K. A., C. Peltier, J. R. Lukos, J. N. Norris, and A. T. Biggs. 2019. *Comparison of Virtual Reality and Augmented Reality: Safety and Effectiveness*. NAMRU-D-19-67. Dayton, OH: Naval Medical Research Unit.

* Piechowski, S., W. Pustowalow, M. Arz, J. Rittweger, E. Mulder, O. T. Wolf, B. Johannes, and J. Jordan. 2020. "Virtual Reality as Training Aid for Manual Spacecraft Docking." *Acta Astronautica* 177: 731–736.

* Pollard, K. A., A. H. Oiknine, B. T. Files, A. M. Sinatra, D. Patton, M. Ericson, J. Thomas, and P. Khooshabeh. 2020. "Level of Immersion Affects Spatial Learning in Virtual Environments: Results of a Three-Condition Within-Subjects Study with Long Intersession Intervals." *Virtual Reality* 24(4): 783–796.

* Pulijala, Y., M. Ma, M. Pears, D. Peebles, and A. Ayoub. 2018. "Effectiveness of Immersive Virtual Reality in Surgical Training—A Randomized Control Trial." *Journal of Oral and Maxillofacial Surgery* 76(5): 1065–1072.

* Ragan, E. D., D. A. Bowman, R. Kopper, C. Stinson, S. Scerbo, and R. P. McMahan. 2015. "Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task." *IEEE Transactions on Visualization and Computer Graphics* 21(7): 794–807.

* Reed, D., C. Maraj, J. Hurter, and L. Eifert. 2019. "Simulations to Train Buried Explosives Detection: A Pilot Investigation." In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

* Reiner, A. J., H. M. Vasquez, G. A. Jamieson, and J. G. Hollands. 2022. "Comparing an Augmented Reality Navigation Display to an Electronic Map for Military Reconnaissance." *Ergonomics* 65(1): 78–90.

* Ricca, A., A. Chellali, and S. Otmane. 2021. "Comparing Touch-based and Head-Tracking Navigation Techniques in a Virtual Reality Biopsy Simulator." *Virtual Reality* 25(1): 191–208.

* Siu, K. C., B. J. Best, J. W. Kim, D. Oleynikov, and F. E. Ritter. 2016. "Adaptive Virtual Reality Training to Optimize Military Medical Skills Acquisition and Retention." *Military Medicine*, 181(suppl_5): 214–220.

Smith, S. P., and S. Du'Mont. 2009. "Measuring the Effect of Gaming Experience on Virtual Environment Navigation Tasks." In *2009 IEEE Symposium on 3D User Interfaces,* 3–10. https://doi.org/10.1109/3DUI.2009.4811198.

* Song, H., T. Kim, J. Kim, D. Ahn, and Y. Kang. 2021. "Effectiveness of VR Crane Training with Head-Mounted Display: Double Mediation of Presence and Perceived Usefulness." *Automation in Construction*, 122, 103506.

Sutherland, I. E. 1968. "A Head-Mounted Three Dimensional Display." In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I* (pp. 757–764).

* Urbano, D., M. de Fátima Chouzal, and M. T. Restivo. 2020. “Evaluating an Online Augmented Reality Puzzle for DC Circuits: Students' Feedback and Conceptual Knowledge Gain.” *Computer Applications in Engineering Education* 28(5): 1355–1368.

* Wallgrün, J. O., M. M. Bagher, P. Sajjadi, and A. Klippel. 2020. “A Comparison of Visual Attention Guiding Approaches for 360 Image-based VR Tours.” In *Proceedings for 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 83–91.

* Wang, C.-H., N.-H. Tsai, J.-M. Lu, and M.-J. J. Wang. 2019. “Usability Evaluation of an Instructional Application Based on Google Glass for Mobile Phone Disassembly Tasks.” *Applied Ergonomics* 77: 58–69.

* Webel, S., U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche. 2013. “An Augmented Reality Training Platform for Assembly and Maintenance Skills.” *Robotics and Autonomous Systems* 61(4): 398–403.

* Westerfield, G., A. Mitrovic, and M. Billinghurst. 2013. “Intelligent Augmented Reality Training for Assembly Tasks.” In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik (eds)*International Conference on Artificial Intelligence in Education,* 542–551. Springer, Berlin, Heidelberg.

* Whitmer, D. E., D. Ullman, and C. I. Johnson. 2019. “Virtual Reality Training Improves Real-World Performance on a Speeded Task.” In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63(1): 1218–1222. https://doi.org/10.1177/1071181319631013.

* Winther, F., L. Ravindran, K. P. Svendsen, and T. Feuchtner. 2020. “Design and Evaluation of a VR Training Simulation for Pump Maintenance.” In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–8. https://doi.org/10.1145/3334480.3375213

* Yang, Z., J. Shi, W. Jiang, Y. Sui, Y. Wu, S. Ma, and H. Li. 2019. “Influences of Augmented Reality Assistance on Performance and Cognitive Loads in Different Stages of Assembly Task.” *Frontiers in Psychology* 10: 1703.

Yilmaz, Z. A., and V. Batdi. 2021. “Meta-Analysis of the Use of Augmented Reality Applications in Science Teaching.” *Journal of Science Learning*, 4(3), 267–274.

* Yoganathan, S., D. A. Finch, E. Parkin, and J. Pollard. 2018. “360° Virtual Reality Video for the Acquisition of Knot Tying Skills: A Randomised Controlled Trial.” *International Journal of Surgery* 54: 24–27.

Abbreviations

AR	augmented reality
COTS	commercial off-the-shelf
DoD	Department of Defense
HCI	Human-Computer Interaction
HMD	Head-Mounted Display
HUD	heads-up display
MitS	mirror in the sky
MR	mixed reality
NASA	National Aeronautics and Space Administration
PSSUQ	Post-Study System Usability Questionnaire
RDBMS	relational database management system
SUS	System Usability Scale
TLX	task load index
VR	virtual reality
XR	extended reality

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* December 2022	**2. REPORT TYPE** FINAL	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE Training Effectiveness Framework for Augmented and Virtual Reality: Developing a Knowledge Base	**5a. CONTRACT NUMBER** HQ0034-19-D-0001
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Belanich, James Moses, Franklin L. Fedele, Emily A.	**5d. PROJECT NUMBER** AK-2-4836
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses Systems and Analyses Center 730 East Glebe Road Alexandra, VA 22305-3086	**8. PERFORMING ORGANIZATION REPORT NUMBER** IDA Document D-33349
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory Airman Systems Directorate Wright-Patterson AFB, OH 45433	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A. Approved for public release; distribution is unlimited (18 May 2023).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Numerous meta-analyses show that augmented reality (AR) and virtual reality (VR) technologies can be effective for training and performance enhancement. However, the broad assertion of AR/VR effectiveness is overly simplistic and clouds important factors influencing the generalizability of findings. Studies of AR/VR effectiveness come from disparate disciplines and perspectives that may complicate the applicability of findings from effectiveness studies. The studies include disciplines such as computer science and engineering focused on technology, education and training focused on instructional methods, and other disciplines (e.g., medical, military, and construction) focused on their specializations. A framework, developed for this project, cuts across differences by employing a common language with four dimensions and an outcome category, captured through five questions: 1) What is the technology? 2) What are the tasks/skills being trained? 3) Who are the users? 4) How is the technology used or integrated or trained in a situation/course? 5) What was the study outcome? That framework provides the foundation for an online knowledge base with characteristics that users can query to help probe the evidence for AR/VR technologies and their effectiveness. This report is about the framework's development, its application to a searchable knowledge base, and the kinds of results that enable users to organize and summarize AR/VR effectiveness study findings for their own needs. An initial knowledge base of 64 studies that employs the framework can grow to provide an ever-improving resource for users to understand how AR/VR can be effective.

15. SUBJECT TERMS

Augmented Reality (AR); Virtual Reality (VR)

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Bennett, Winston
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			
Unclassified	Unclassified	Unclassified	Same as Report	65	**19b. TELEPHONE NUMBER** *(Include area code)* 937.938.2550

Reset