NISTIR 8422

Augmented Reality (AR) Usability Evaluation Framework:

The Case of Public Safety Communications Research

Yee-Yin Choong
Kurtis Goad
Kevin C. Mangold

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8422

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8422

Augmented Reality (AR) Usability Evaluation Framework:

The Case of Public Safety Communications Research

Yee-Yin Choong
Kurtis Goad
Kevin C. Mangold
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8422

April 2022

U.S. Department of Commerce
Gina M.Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology Interagency or Internal Report 8422
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8422, 24 pages (April 2022)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8422

Abstract

Augmented Reality (AR) is an enhanced version of reality created by the use of technology to overlay digital information on an image of something being viewed through a device. AR solutions have potential uses in many fields such as education, healthcare, retail, repair/maintenance, manufacturing, and gaming.

Any well-conducted and well-planned product development project should follow an iterative human-centered process. Throughout the development lifecycle, usability evaluations with target users should be conducted to ensure that the product can be used by the specified users to achieve the specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. This AR Usability Evaluation Framework provides guidance on planning user-based usability evaluations of AR technology. While this report demonstrates the framework using a firefighting scenario in the public safety domain, the framework is applicable and can be expanded to other domains where user-based AR usability evaluations will be performed.

Applying the framework within AR solution development lifecycles will provide the following benefits:

- Creating explicit structures for user-based evaluations
- Providing a consistent terminology and an initial set of usability metrics
- Facilitating comparability across AR research and development efforts
- Facilitating sharing of usability evaluation results
- Facilitating establishing human-centered AR design guidelines

This report provides a five-component AR Usability Evaluation Framework to facilitate systematic planning of usability evaluations to ensure successful evaluations and collection of useful usability data for product improvement. The five components are: (1) Determine evaluation scope; (2) Identify users and context of use; (3) Develop evaluation scenario and tasks; (4) Select applicable usability metrics; and (5) Define usability measures for selected metrics. Following this framework to conduct usability evaluations throughout development cycle will help reduce development cost and bring the AR solutions to market faster, while providing usable products that are easy, quick, comfortable, and safe to use.

Key words

Augmented Reality; Usability; Usability Evaluation; Human-centered design; User needs and requirements; Human Factors and Ergonomics; Public safety communications research.

Audience

This report is primarily intended for designers, developers, vendors, and researchers of augmented reality technology.

Disclaimer

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National

Institute of Standards and Technology, nor does it imply that the products mentioned are necessarily the best available for the purpose.

Table of Contents

1. INTRODUCTION 4

2. AR USABILITY EVALUATION FRAMEWORK 5

2.1. The Importance of Usability Evaluation 5
2.2. Five Components of the AR Usability Evaluation Framework 6
2.2.1. Determine Evaluation Scope 6
2.2.2. Identify Users and Context of Use 6
2.2.3. Develop Evaluation Scenario and Tasks 6
2.2.4. Select Applicable Usability Metrics 7
2.2.4.1. Performance-based metrics 7
2.2.4.2. Behavioral and Physiological based metrics 8
2.2.4.3. Self-reported metrics 9
2.2.4.4. Issues-based metrics 9
2.2.5. Define Usability Measures for Selected Metrics 9

3. APPLYING THE AR USABILITY EVALUATION FRAMEWORK – A FIREFIGHTING USE CASE 10

3.1. Evaluation Scope – Obstacle Detection and Navigation 10
3.2. Users and Context of Use – Firefighting 10
3.3. Scenario and Tasks – Search in Apartment Fire 10
3.3.1. The Firefighting Scenario 11
3.3.2. Search Process Chosen for AR Usability Evaluations 12
3.4. Usability Metrics Selection 13
3.5. Usability Measures 15
3.5.1. Data Sources 17
3.6. Evaluation Setup 19

4. CONCLUSION 19

REFERENCES 20

APPENDIX A: APARTMENT FIRE SCENARIO – COMPLETE TASK ANALYSIS FLOWCHART 21

List of Tables

Table 1 Performance Metrics for AR Usability Evaluation 7
Table 2 Behavioral and Physiological Metrics for AR Usability Evaluation 8
Table 3 Self-reported Metrics for AR Usability Evaluation 9
Table 4 Issues-based Metrics for AR Usability Evaluation 9
Table 5 Examples of Metrics Chosen for Search Process Usability Evaluation 14
Table 6 Examples of Measures Chosen for Search Process Usability Evaluation 16
Table 7 Example of Data Sources for Search Process Usability Evaluation 18

List of Figures

Figure 1 Five Components of the AR Usability Evaluation Framework 6
Figure 2 Search and Rescue Processes 12

1. Introduction

Augmented Reality (AR), which dates back to 1968 [1][2], is "an enhanced version of reality created by the use of technology to overlay digital information on an image of something being viewed through a device (such as a smartphone camera)" [3] or head-mounted display (HMD). Over more than 50 years of advancements, the equipment required to use AR went from taking up entire rooms to small headsets that can weigh less than two pounds.

Not to be confused with Virtual Reality (VR), where the user does not generally see through to the real world, AR users are still able to see some of reality. Whether the AR device is a headset (or glasses), smartphone, or tablet, they all have a viewer. Generally, AR viewers come in two forms:

- a transparent lens, or
- a digital display such as a smartphone, OLED (organic light-emitting diode), and LCD (liquid crystal display)

For the transparent lens, digital enhancements are accomplished by projecting targeted light onto the lens, thus serving as an overlay to the real-world view. When the device is using a digital display, the user is unable to see through the screen because of the display's internal components. To provide the user with the real-world view, an integrated camera is used to capture the real-world view and relay it to the user through the digital display, i.e., displaying a video feed. Digital enhancements are made to the video feed before being displayed to the user.

AR has many potential uses. It can be and is leveraged in many fields, to name a few: education, healthcare, retail, repair/maintenance, manufacturing, and gaming. It has the power to train new employees or students, allow consumers to customize features of a product they are considering purchasing and view it in real time, assist technicians when diagnosing equipment issues or building new equipment, and provide entertainment to users.

The public safety domain is also a field that could benefit from AR technologies. For example, a firefighter could have vital statistics displayed inside their SCBA (self-contained breathing apparatus) mask instead of grabbing for multiple sensors or gauges, or even have a building schematic shown with their current location, assisting them in navigating a low-visibility environment. A law enforcement officer at a routine traffic stop could scan a driver's license and pull up information about the individual without leaving them unsupervised. An Emergency Medical Technician (EMT) could display a patient's vitals or medical history, so they are able to provide adequate treatment en route to a nearby hospital.

AR has the potential to make significant impacts that assist first responders in their daily responsibilities. However, any solution would need to be developed in such a way that it is easy to use, intuitive, and does not impede the user in any way. This framework intends to lay out guidance of how to plan and assess the usability of an AR solution.

2. AR Usability Evaluation Framework

Usability is defined by the International Organization for Standardization (ISO: 9241-11) as "*the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" [4]. According to ISO:9241-11, effectiveness is related to accuracy and completeness of achieving specified goals; efficiency deals with resources in relation to the results achieved such as time, human effort, cost and materials; and satisfaction has to do with the extent to which the user's physical, cognitive, and emotional responses that result from the use of a system, product or service meet the user's needs and expectations [4].

2.1. The Importance of Usability Evaluation

A well-conducted and well-planned product development project should follow an iterative two-stage human-centered process [5]. At the first stage, a human factors and ergonomics (HFE) professional(s) should be included ensure that all good HFE principles are considered and applied from the beginning design stage throughout the entire product development lifecycle. The second stage starts when a prototype has been designed and developed, to go through systematic, science-based, and data-driven usability evaluations performed by trained usability professionals. Comprehensive usability evaluations should involve target users with representative tasks in realistic operational environments. The two stages can overlap and be iterative as modifications may be necessary based on the outcomes of the usability evaluations, which will require the product to go back to stage one.

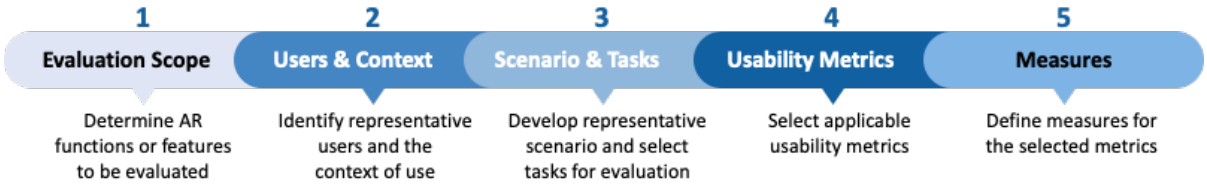
Usability evaluation is an iterative process which frequently results in needed product modifications before production and deployment to ensure user requirements are met for performance excellence. Too often, product development teams consider evaluation to be the final stage when the system is nearing completion. One major risk of delaying the evaluation to only toward the end of product development is that the evaluation may reveal major issues with the system that are too costly or too difficult to correct in the latter stages of development. By incorporating feedback from representative users throughout the design of a system, it is easier to identify major problems in a system at a much earlier stage.

Some of the widely used usability evaluation methods include but are not limited to the following: contextual inquiries/naturalistic observation, structured questionnaires, individual/group discussions on performance and usability issues, cognitive walkthroughs (a usability inspection method), heuristic evaluations (a holistic view to catch usability problems), and user-based usability evaluations [6].

In this report, we focus on user-based usability evaluations. These involve real-world, representative users 'trying out' or testing a design of a system, while user experience data are collected, analyzed, and documented to identify areas for improvements where the system does not meet users' expectation, or does not support users performing representative tasks. Areas that work well for users are also documented. Usability evaluation sessions are often recorded and/or observed by members of the product team to identify usability issues with the system. The goal of a usability evaluation is to assess the technology, not to test the user.

2.2. Five Components of the AR Usability Evaluation Framework

This AR Usability Evaluation Framework provides a common language and consistent process for planning user-based usability evaluations of AR technology. There are five components in planning a user-based AR usability evaluation as shown in Figure 1.



2.2.1. Determine Evaluation Scope

Usability evaluations should be performed by trained usability professionals, who will be referred to as Test Admin in the remainder of this report. When planning a usability evaluation, it may not always be feasible to test the entire system in one usability evaluation. So, the first step is for the Test Admin, working closely with the product development team, to determine the evaluation objectives and scope – which functions or features in the AR system are to be evaluated.

2.2.2. Identify Users and Context of Use

A well-planned product development project would follow the two-stage human-centered process, where target users and the context of use are well identified and defined in the first stage [5][6]. When planning a usability evaluation, it often involves testing a subset of the system functions, as described in section 2.2.1. Thus, it is important to identify the primary users of the selected functions and the context where those functions will be used.

2.2.3. Develop Evaluation Scenario and Tasks

With the evaluation objectives and scope, primary users and context of use determined, the next step is to develop a representative scenario and select relevant user tasks for evaluating the selected AR functions. To help determine the tasks within the scenario, a task analysis should be conducted. Task analyses define the tasks to be completed in a given scenario based on the goals of the users in that scenario [7]. Task analyses are useful for narrowing the scope of evaluation as well as choosing which tasks are appropriate for assessing the selected functions. In this framework, we will focus on four task types that users may perform in an AR usability evaluation session: Action, Communication, Detection, and Monitoring tasks.

- **Action task (A)** – the user performs a specific action to complete the task or make change(s) in the system or physical environment
- **Communication task (C)** – the user is required to communicate with another user or the Test Admin
- **Detection task (D)** – the user needs to detect targets (objects or information) placed in the environment

- **Monitoring task (M)** – the user must track specific states, conditions, objects, or people over time and could be required to monitor multiple targets simultaneously.

2.2.4. Select Applicable Usability Metrics

Following a systematic, science-based, and data-driven methodology, the objective of usability evaluation is to measure the usability of a technology against a set of well-defined usability metrics for assessing the user experience interacting with the technology. In this section, a list of usability metrics for evaluating AR technology is provided. From this pool of usability metrics, the Test Admin will select applicable metrics for the developed scenario and user tasks, as described in 2.2.3, to evaluate the usability of the selected AR functions.

The usability metrics will be described in detail and are organized with the following components:

- **Metrics Categories:** Performance, Behavioral and Physiological, Self-reported, and Issues-based [8]
- **Usability Dimensions:** which usability dimension a metric falls under – Effectiveness, Efficiency, or Satisfaction
- **Usability Metrics and Description:** each usability metric and its associated description
- **Data Types:** can be quantitative such as counts, ratios, time, scale ratings; and/or qualitative such as observations, open-ended responses
- **Actors:** two types of actors, *Users* who participate in usability evaluation; and *Test Admin* who is responsible to plan and conduct the usability evaluation, and facilitate collection of evaluation data

2.2.4.1. Performance-based metrics

Performance-based metrics are metrics directly related to the extent to which a participating user can successfully accomplish the target scenario and tasks within a reasonable timeframe. Performance metrics capture how participants complete target tasks and respond to planned questions during a usability evaluation and are among the best ways to evaluate the effectiveness and efficiency of products. There are opportunities for product improvement if users make many errors or if users take a much longer time to complete a task than what was expected. Table 1 lists performance-based metrics for evaluating AR technology.

Table 1 Performance Metrics for AR Usability Evaluation

Usability Dimension	Usability Metrics	Description	Data Type	Actor(s)
Effectiveness	Task completion	Whether or not the user completes the task intended	Binary success or Levels of success	User Test Admin
	Session completion	Whether or not the user completes the usability session	Binary success or levels of success	User Test Admin
	Completeness	Ratio of events completed to total events expected	Ratio	User Test Admin
	Accuracy/Errors	Frequency of user events that do (or do not) cause an expected outcome	Counts	User Test Admin
	Spatial Accuracy	Correct interactions with real or virtual object intended, not missing contact with the objects	Counts or ratio	User Test Admin
	Event deviation	Events performed by the user that do not aid in completing the task intended	Counts or ratio	User Test Admin

Efficiency	Time-on-Task	Time spent performing a task	Task duration	User Test Admin
	Time until Event	The time between a predefined stimuli presentation and the start of a user event	Task duration	User Test Admin
	Time-on-Session	Time spent performing the usability session	Session duration	User Test Admin
Effectiveness ***Efficiency***	Learnability	Whether/how user's performance differs (improves or degrades) over time	Multiple trials over time	User Test Admin

2.2.4.2. Behavioral and Physiological based metrics

Different from performance-based, behavioral and physiological based metrics are metrics related to participating users' behaviors and emotions demonstrated during a usability evaluation. Users may smile, laugh, frown, grimace, or fidget. They may show a wide range of emotions such as stress, excitement, frustration, and surprise. They may stare at certain objects or look around aimlessly. These behavioral and physiological metrics provide valuable insights into the user's experience interacting with the technology being evaluated. Table 2 lists behavioral and physiological based metrics for evaluating AR technology.

Table 2 Behavioral and Physiological Metrics for AR Usability Evaluation

Usability Dimension	Usability Metrics	Description	Data Type	Actor(s)
Effectiveness	Eye Tracking–Scan patterns	The order or pattern in which the user looks at while completing a task	Eye-tracking heat maps	User Test Admin
	Mental workload	An index to assess self-reported mental workload for a task or the session, reported by the user	NASA-TLX[1]	User Test Admin
Efficiency	Eye Tracking–Dwell time	Duration of eye gaze directed at a specific target	Duration	User Test Admin
	Eye Tracking–number of fixations	Frequency of instances of eye gazes directed at a specific location or object	Counts	User Test Admin
	Communication effort–Speaker turns	Number of turns in conversation between two speakers	Counts	User Test Admin
	Communication Effort–Words spoken	Number of words spoken by one user	Counts	User Test Admin
	Communication Effort–Grounding questions asked	Number of questions asked to another user or Test Admin in order to help understand information presented by the system or in the environment of the user	Counts	User Test Admin
Effectiveness ***Efficiency***	Verbal	User's verbal interactions with Test Admin during the session	Observations	User Test Admin
	Nonverbal	User's nonverbal information observed by Test Admin during the session	Observations	User Test Admin
	Facial expressions	User's facial expressions observed by Test Admin during the session	Observations	User Test Admin

[1] NASA Task Load Index (TLX) is a subjective workload assessment tool which allows users to perform subjective workload assessments on operator(s) working with various human-machine interface systems (https://humansystems.arc.nasa.gov/groups/tlx/).

2.2.4.3. Self-reported metrics

Self-reported metrics are data gathered directly from the participating users, usually in the form of quantitative questionnaire with scale ratings or qualitative responses such as open-ended responses or interviews. If the self-reported metrics are taken prior to interacting with the technology being evaluated, the user can answer questions regarding their expectations of the technology or past experiences with similar technology for comparison. If the self-reported metrics are taken after the usability evaluation, the user can answer questions regarding their perceptions of the technology and their experience interacting with the technology. Table 3 lists self-reported metrics for evaluating AR technology.

Table 3 Self-reported Metrics for AR Usability Evaluation

Usability Dimension	Usability Metrics	Description	Data Type	Actor(s)
Satisfaction	Pre-Session Expectations	An index of questions answered by the user, before using the system to assess the user's expectations about the system prior to using it	Scale ratings and/or open-ended	User Test Admin
	Post-task Post-session	Questions can include: • Ease of Use • Task and Content Specific Questions • Perception of Outcomes/Interactions • Comfort • Learnability	Scale ratings and/or open-ended	User Test Admin

2.2.4.4. Issues-based metrics

Issues-based metrics are usability issues identified with severity ratings assigned by the Test Admin. During a usability evaluation, the Test Admin notes areas of concern or user confusion and may also ask the user to 'think aloud' in order to better understand why a user is behaving in a certain manner. Not all usability issues are the same–some may mildly annoy or frustrate users while others can cause them to make the wrong decisions or lose data. Severity ratings assigned to usability issues help product development team prioritize and focus their attention on the issues that really need to be addressed to improve user experience and system performance. Severity ratings can be assigned by the Test Admin during or after a usability evaluation. Table 4 lists issues-based metrics for evaluating AR technology.

Table 4 Issues-based Metrics for AR Usability Evaluation

Usability Dimension	Usability Metrics	Description	Data Type	Actor
Effectiveness ***Efficiency***	Identify issues and assign severity ratings	Usability issues identified by the Test Admin during the session	Counts and severity ratings	Test Admin

2.2.5. Define Usability Measures for Selected Metrics

Once the applicable metrics are selected for the developed scenario and user tasks for evaluating the selected AR functions, the Test Admin will define how each metric selected will be measured for the AR usability evaluation. For example, timing metrics such as Time-on-Task can be measured by calculating elapsed time of a user event from the time-stamp data in the system logs. We will use a Firefighting scenario in Section 3 to demonstrate the end-to-end process of applying the AR Usability Evaluation Framework.

3. Applying the AR Usability Evaluation Framework – A Firefighting Use Case

This section uses a hypothetical use case to illustrate how each step of the five-component AR Usability Evaluation Framework could be applied to plan the evaluation of an AR solution designed for first responders. Specifically, the use case being evaluated is designed for firefighting. Firefighting scenarios demonstrate the complex and dynamic nature of first responder situations in that fire incidents typically involve several team members performing multiple different tasks simultaneously. The AR solution chosen for our demonstration is a Heads-Up Display (HUD) that can display information and offer assistance relevant to firefighters for maintaining safety and completing their mission. The HUD is designed to aid in both Incident Command (IC) and responding firefighters perspectives. The IC perspective may display visual information on interactive maps projected into real space such as birds eye views, team status, and resource availability. The responding firefighters perspective may include functions such as obstacle detection and navigation and show information helpful to the individual firefighters while completing their mission such as room temperature or personal air level.

3.1. Evaluation Scope – Obstacle Detection and Navigation

Usability evaluations should be iterative throughout the product development cycle. For each usability evaluation, a logical subset of the system functions should be chosen containing representative tasks. The evaluation scope of our example focuses on the responding firefighters perspective to evaluate the AR solution's obstacle detection and navigation functions designed to aid firefighters performing essential tasks in a fire incident.

3.2. Users and Context of Use – Firefighting

Determining the evaluation scope helps the Test Admin and the product team further define the context of use and representative users to recruit for the evaluation. Specifically, firefighters who have experience in responding to fire scenes are target participants based on the evaluation scope and functions chosen for evaluation–obstacle detection and navigation. Target users must be consulted, here responding firefighters, to establish appropriate context for conducting the evaluation. The context of use includes using the AR solution in a dangerous, high-risk environment, such as a burning building, and may include life or death situations, making it imperative that whatever solution is provided to firefighters be as reliable as possible. Firefighters also work in a multitask-based and team-oriented environment in which they are responsible for several ongoing tasks coordinated between multiple team members. Therefore, the test environment of the proposed AR solution must, within reason, replicate these conditions closely for the evaluation to be accurate, but without causing real harm to the participants.

3.3. Scenario and Tasks – Search in Apartment Fire

The Test Admin will work closely with the product team to develop a scenario and representative tasks for the usability evaluation through consultation with subject matter experts, i.e., firefighters in this use case. The scenario developed for the usability evaluation will incorporate the previously established context of use. To demonstrate how this step

works, we consulted firefighters and developed a narrative of an apartment fire to 'set the stage' for what tasks will be involved in the scenario.

> **Narrative:** A fire was reported to have started in an apartment in a building. The local fire department was alerted and dispatched, a crew donned their gear and loaded onto a fire engine and a ladder truck, and they drove to the scene following a route planned by the command officer. Upon arrival, smoke can be seen coming from an open apartment window on the 2nd floor of the two-story building. Other building residents reported to the command officer that they did not know whether the resident(s) of the apartment on fire was in the building or not. The fire engine was positioned next to the fire hydrant outside the building.

3.3.1. The Firefighting Scenario

In the firefighting scenario: 1) tasks will be completed by a team of firefighters; 2) firefighting will be a dynamic process in which decisions must be made and questions answered that may change the course of action of each team member and may change how and which tasks will be completed. To demonstrate, we conducted a task analysis of fire incident response to the apartment fire narrative described above in Section 3.3. The task analysis helped determine what the essential tasks, as well as the decisions associated with each task, would most likely be during the apartment firefighting scenario. The task analysis was iterative in nature and involved consulting a firefighter, incorporating the feedback into multiple iterations of the analysis, and compiling the results into a visual flowchart (see Appendix A for the complete task analysis flowchart).

The flowchart shows activities and decision points in each step and the flow in completing the mission. In addition, badges containing some or all of the letters–A, C, D, and M–are placed in the top right corner of some of the component boxes. These represent which of the four task types (Action, Communication, Detection, and Monitor) are involved in each component when using an AR solution. Figure 2 shows the portion of two sub-processes of the analysis–Search and Rescue.

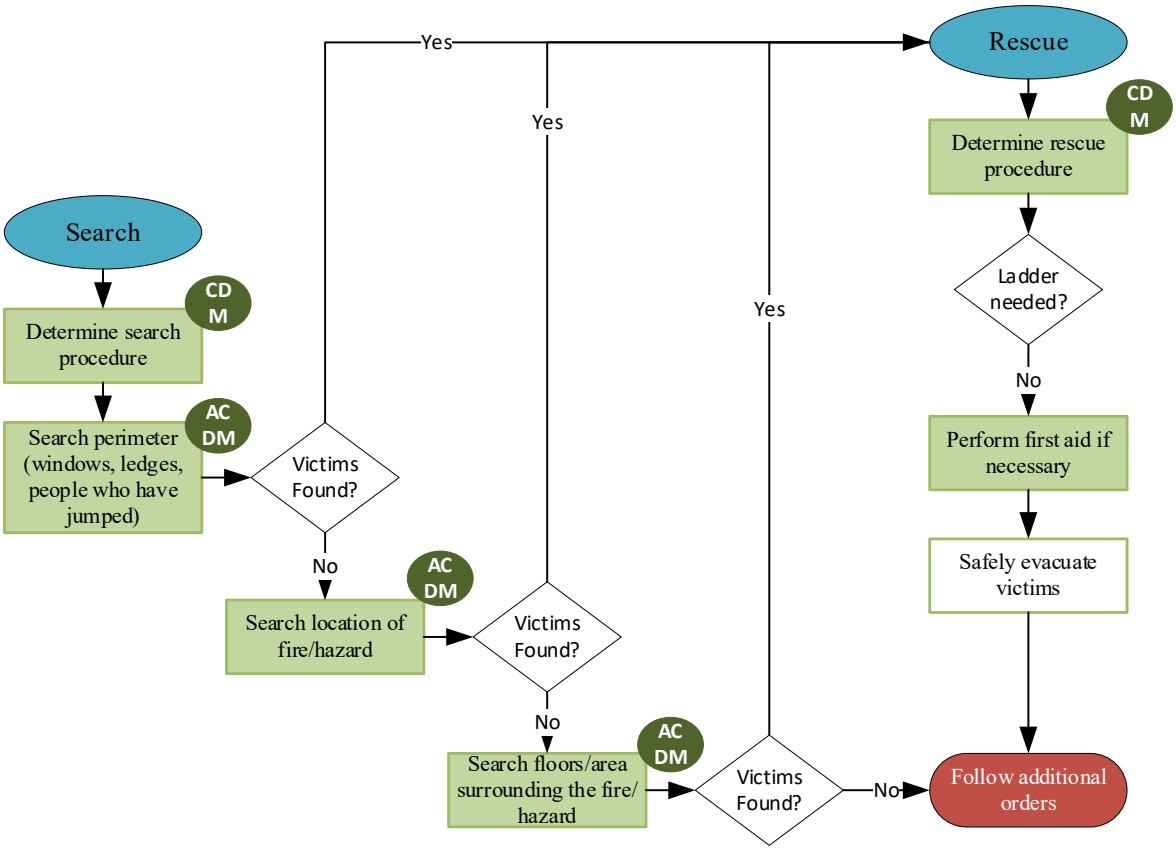


Figure 2 Search and Rescue Processes

3.3.2. Search Process Chosen for AR Usability Evaluations

A well conducted task analysis will often demonstrate that there are many tasks that could be used during an evaluation. However, from all possible choices, only representative tasks in a given scenario will be selected to align with the evaluation scope and match the specific functions being evaluated, i.e., Obstacle Detection and Navigation in our example. Using the apartment fire use case to demonstrate how this step of choosing representative tasks is done, we chose the search process as it contains all four task types and the functions being evaluated (obstacle detection and navigation). The user will be introduced to each task with a statement either pre-configured in the AR headset or given by the Test Admin.

Examples of the four specific tasks chosen are as follows:

- **Action Task**: *Search the location of the fire for victims*.
  - **Description:** In this task, the user will physically walk through the test environment searching for any victims (e.g., mannequins) that may be present within predetermined rooms or locations.
  - **Task Statement:** The task will be introduced with a statement such as: "you walk into the apartment; your task is to determine if there are any victims and identify their locations within the apartment."

    - **Goal:** This task is to evaluate whether the chosen AR functions, especially the navigation function, improve completion of the essential task of searching for victims.
- **Communication Task**: *Report conditions over the radio*.
    - **Description**: Here the user will be asked to report certain conditions back to a commander (e.g., Test Admin) over a radio as the conditions occur.
    - **Task Statement**: The user will be introduced to this task with a statement such as: "As you navigate through the building you will be reporting to your commanding officer, using your radio, the location of the fire and if there are any victims present."
    - **Goal**: The goal of this task is to evaluate whether the AR system facilitates faster and more effective communication between the user and other team members.
- **Detection Task**: *Detect hazards*.
    - **Description**: The user will be responsible for detecting certain preset hazards that occur during task completion.
    - **Task Statement**: The user will be introduced to this task with a statement such as: "The objective is to find a clear path to navigate through the building while searching the apartment, do your best to avoid any obstacles."
    - **Goal**: The goal of this task is to evaluate whether the chosen functions, especially the obstacle detection function, improve the user's ability to detect hazards and help identify a clear path.
- **Monitor Task**: *Monitor your air level*.
    - **Description**: The user's air level will change a certain number of times throughout the scenario and the user will be responsible for monitoring these changes.
    - **Task Statement**: The user will be introduced to this task with a statement such as: "Your air level will be displayed in your mask. Report to the Test Admin any changes while searching the apartment."
    - **Goal**: The goal of this task is to evaluate the way the AR solution displays important data to the user and whether it helps the user monitor information status.

3.4. Usability Metrics Selection

Following the selection of representative tasks, applicable usability metrics will be selected by the Test Admin as mentioned in Section 2.2.4. These metrics are carefully chosen according to the evaluation goals determined in collaboration with the product team in order to define how usability of the AR solution will be assessed. For each representative task in the apartment fire example, several usability metrics were chosen from the larger list of metrics described in Section 2.2.4, organized by metric category and usability dimension. For demonstration purpose on how to select appropriate usability metrics, we focused on two metric categories: performance-based metrics; behavioral and physiological based metrics. Table 5 shows examples of metrics chosen for tasks in the search process usability evaluation.

Performance-based metrics

Some common performance usability metrics, measuring both effectiveness and efficiency, were selected for the representative tasks such as task completion, task completeness, time on task, and errors. These metrics are common to usability evaluations and can easily be applied to AR solutions. They are included because they indicate whether or not the AR function can help the user succeed in completing their mission, and if it is able to do so within a reasonable or expected amount of time.

Table 5 Examples of Metrics Chosen for Search Process Usability Evaluation

Task	Metric Category	Usability Dimension	Metrics
Search location of hazard	Performance	Effectiveness	• Task Completion • Task Completeness
		Efficiency	• Time on Task
	Behavioral Physiological	Effectiveness	• Event Deviation • Spatial Accuracy
		Efficiency	• Eye Tracking (Dwell Time) • Eye Tracking (Number of Fixations)
Report conditions over radio	Performance	Effectiveness	• Task Completion • Task Completeness • Errors
		Efficiency	• Time on Task
	Behavioral Physiological	Effectiveness	• Event Deviation
		Efficiency	• Speaker Turns • Words Spoken • Grounding Questions Asked
Detect hazards	Performance	Effectiveness	• Task Completion • Task Completeness • Errors
		Efficiency	• Time Until Event
	Behavioral Physiological	Effectiveness	• Eye Tracking (Sequence/Scan Patterns)
		Efficiency	• Eye Tracking (Dwell Time) • Eye Tracking (Number of Fixations)
Monitor air level	Performance	Effectiveness	• Task Completion • Task Completeness • Errors
		Efficiency	• Time Until Event
	Behavioral Physiological	Effectiveness	• Eye Tracking (Sequence/Scan Patterns)
		Efficiency	• Eye Tracking (Dwell Time) • Eye Tracking (Number of Fixations)

Behavioral and Physiological based metrics

In the apartment fire example, Behavioral and Physiological metrics were chosen more specifically for their associated representative tasks to assess effectiveness and efficiency.

For example, to assess the AR solution's effectiveness on the task "Search location of fire for victims," two metrics were chosen. The first one was spatial accuracy–measuring the

accuracy with which the user interacts with the real and virtual aspects of the environment while using the AR HUD. While wearing a HUD that shows virtual objects in a real physical space, it is crucial that the user is able to interact with the environment without interference caused by the HUD. Measuring spatial accuracy indicates whether there is any interference to the user's interaction within the environment. Another effectiveness metric chosen was event deviation. This metric allows the Test Admin to account for any behaviors that deviate unexpectedly and do not aid in completing the task. For example, if the user needs to adjust the HUD or performs any unnecessary navigation of the features on the interface while trying to perform a task.

Eye tracking metrics are particularly useful for assessing effectiveness and/or efficiency of AR solutions, as the solutions directly affect the users' sight. For example, the sequence/scan pattern metric was chosen to assess effectiveness in the "Detect hazards" task and "Monitor air level" task. In the "Detect hazards" task, the sequence/scan pattern metric will measure whether the user follows the optimal path when detecting hazards during the scenario. This will help determine the effectiveness of the AR system's ability to alert the user of potential hazards. Similarly, in the "Monitor air level" task, the sequence/scan pattern metric will measure how the user looks at the air level information and detects changes.

As mentioned earlier, efficiency metrics often include time as an important factor in measurement. AR solutions with great efficiency will reduce time to complete a task. Eye tracking dwell time and number of fixations can be used to measure efficiency applicable to three of our example tasks: "Search location of fire for victims," "Detect hazards," and "Monitor air level." Dwell time will measure how long the user looks at specific targets when completing the task and may indicate whether the HUD is efficient at helping the user perform the tasks. The number of fixations will measure how many times the user looks at a target object, indicating whether the AR function being evaluated is efficient at helping users complete their goals. For example, if the user has to look at an alert more than once to understand the information being shown it may be determined to be inefficient, depending on the goal of the system function.

For the communication task, "Report conditions over radio," different types of efficiency metrics were chosen to measure communication effort: speaker turns, words spoken, and grounding questions asked. The HUD is designed to facilitate easier communication between users. The chosen communication metrics measure efficiency by seeing how much the user is speaking. Ideally, the more efficient the HUD is at facilitating communication, the less the user will need to talk, repeat themselves, or ask clarifying questions compared to circumstances in which the HUD is not being used.

3.5. Usability Measures

Metrics will be chosen based on their relevance to help measure the AR solution's usability with respect to each task and task type. This selection along with metrics being divided by metric category and usability dimension will allow the Test Admin to plan and work with the development team for systematically collecting data for selected metrics. The metrics chosen for each task must have specific corresponding measures defined in order to collect data associated with the metrics during usability evaluation. Table 6 shows examples of measures

associated with each of the metrics chosen for the tasks in the apartment fire scenario evaluation. Table 6 specifies exactly what will be measured during a usability evaluation session. Clear and explicit definitions of measures are especially important when some of the same metrics are chosen for multiple tasks.

Table 6 Examples of Measures Chosen for Search Process Usability Evaluation

Task	Measures	Definition
Search location of hazard	**Task Completion**	Binary Yes/No, does the user search all locations?
	Task Completeness	Out of all locations to search, how many does the user search?
	Time on Task	Time in minutes/seconds spent performing a search of a location.
	Event Deviation	Frequency count of any events performed by the user that do not aid in completing the search task (e.g., unnecessary navigation of interface, physically adjusting interface).
	Spatial Accuracy	Ratio of correct interactions (physical movements intending to interact with a real or virtual object in which the user's movements do not deviate from contact with the object) to total interactions during task completion.
	Eye Tracking (Dwell Time)	Time spent in minutes/seconds looking at target objects or locations during a search task.
Report conditions over radio	**Task Completion**	Binary Yes/No, does the user report all pre-determined conditions?
	Task Completeness	Out of all pre-determined conditions, how many does the user report?
	Errors	Frequency of user evoked events in which the user intends to communicate but fails to do so, communicates with the wrong person, etc.
	Time on Task	Time in minutes/seconds spent communicating to report conditions.
	Event Deviation	Frequency count of any events performed by the user while completing the communication task that do not aid in completing the communication task (e.g., unnecessary navigation of interface).
	Speaker Turns	The number of turns in conversation between a user and another actor (e.g., user or Test Admin).
	Words Spoken	The number of words spoken by the user to another actor (e.g., user or Test Admin).
	Grounding Questions Asked	Number and content of questions asked to another actor (e.g., user or Test Admin) in order to help the user understand a piece of information presented by the system or in the environment of the user.
Detect hazards	**Task Completion**	Binary Yes/No, does the user detect all pre-determined target objects/events for detection?
	Task Completeness	Out of all pre-determined target objects/events, how many does the user detect?
	Errors	Frequency of instances in which a user incorrectly detects objects/events (e.g., the user identifies a hazard that is actually a victim).
	Time Until Event	The time between a predefined hazard related stimuli presentation (visual, auditory, etc.) and the detection of the hazard related stimuli.
	Eye Tracking (Sequence/Scan Patterns)	Eye gaze/ heat map of target objects or events the user looks at.
	Eye Tracking (Dwell Time)	Time spent in minutes/seconds looking at target objects or locations presented for detection.
	Eye Tracking (Number of Fixations)	Frequency count of fixations on target objects or locations presented for detection.
Monitor air level	**Task Completion**	Binary Yes/No, does the user monitor all pre-determined changes in air levels?
	Task Completeness	Out of all pre-determined changes in air level, how many does the user monitor?

	Errors	Frequency of instances in which a user incorrectly monitors self-status information (e.g., reading changes in temperature when trying to read changes in air level).
	Eye Tracking (Dwell Time)	Time spent in minutes/seconds looking at target objects or locations presented for detection.
	Eye Tracking (Number of Fixations)	frequency count of fixations on target objects or locations presented for monitoring.

For example, *Task Completeness* was chosen as an effectiveness metric for the action task "Search location of hazard" and the communication task "Report conditions over radio". The measures for the same metric are different. For "Search location of hazard" task, *Task Completeness* was defined as "Out of all locations to search, how many does the user search?" while for the "Report conditions over radio" task, it was defined as "Out of all pre-determined conditions, how many does the user report?" These definitions state exactly what will be measured in each task so that the stated pre-determined areas for searching and conditions for communication can be prepared for testing. Without the definitions, there is no "best case" to assess the user's interaction with the technology. For example, it might be the goal for the development team to know that the user searched three out of five rooms rather than only knowing that they searched three rooms without knowing how many rooms there were available to be searched.

Defining the specific measures is also important when several tasks are being completed simultaneously, so the Test Admin knows which actions or behaviors are associated with which task. For example, *Errors*, as an effectiveness metric for the Communication task "Report conditions over radio" was defined as "Frequency of user evoked events in which the user intends to communicate but fails to do so, communicates with the wrong person, etc." So, if the user is trying to communicate but accidently presses a button that activates infrared vision (heat detection) it may be labeled as an error. However, if the user presses the button to activate infrared vision while trying to scan for hazards it may not be counted as an error. The same behavior, pressing a specific button, will be measured differently according to what task the user is currently trying to perform. In a dynamic scenario, such as the apartment firefighting chosen for evaluation, these distinctions will allow the Test Admin and the product development team to plan for accurate measurement of all metrics included as data are being collected.

3.5.1. Data Sources

Once the metrics have been selected and the measures defined, it is essential to determine the requirements of how the data for each metric will be collected. During a usability evaluation session involving a dynamic scenario, different types of data will need to be collected simultaneously as the user will be completing several tasks at any given time. The requirements and sources for collecting data during the evaluation of an AR solution fall under one or more of four categories:

- **System** data is data that can be recorded by the technology solution itself, such as event timestamp or eye tracking data which are needed for the time on task and all eye tracking metrics.

- **Test Admin** data is collected from the Test Admin during or after the evaluation. For example, a Test Admin must verify whether and how many tasks are completed to score for the task completion/completeness metrics.
- **User Self-Reported** data is collected from the user. For example, user answers questionnaires to assess mental workload and ease of use.
- **Media** data includes any video/audio/screen recordings or recordings of the user view from the AR headset. This type of data is especially important to collect for the communication and spatial accuracy metrics.

Table 7 shows the sources of data associated with each of the metrics chosen for the usability evaluation in the example use case. Although metrics with "User Self-Reported" data were not discussed in the AR solution example, they are included in Table 7 demonstrating a comprehensive list of data sources. If used in a scenario like the firefighting example, the metrics with user self-reported data could be collected pre- and post-session rather than for each specific tasks.

Table 7 Example of Data Sources for Search Process Usability Evaluation

Metric	Source of Data Required			
	System Data	Test Admin	User Self-Reported	Media
Task/Session Completion		Admin Determination		Video/Audio Recording; Screen/User View Recording
Task/Session Completeness		Admin Determination		Video/Audio Recording; Screen/User View Recording
Error	Button Press Gesture Recognition	Inspection of System and Media Data		Video/Audio Recording; Screen/User View Recording
Time on Task	Time Stamp	Inspection of System and Media Data		Video/Audio Recording; Screen/User View Recording
Session Duration	Time Stamp			
Time Until Event	Time Stamp	Inspection of System and Media Data		Video/Audio Recording; Screen/User View Recording
Eye Tracking (Sequence/Scan Patterns)	Eye Tacking Software	Inspection of User View and Eye Tracking Data		Video/Audio Recording; Screen/User View Recording
Event Deviation	Button Press Gesture Recognition	Inspection of System and Media Data		Video/Audio Recording; Screen/User View Recording
Spatial Accuracy		Visual Inspection of Video and User View Data		Video/Audio Recording; Screen/User View Recording
Eye Tracking (Dwell Time)	Eye Tacking (Time Stamp)	Inspection of User View and Eye Tracking Data		Video/Audio Recording; Screen/User View Recording
Eye Tracking (Number of Fixations)	Eye Tacking Software	Inspection of User View and Eye Tracking Data		Video/Audio Recording; Screen/User View Recording
Speaker Turns		Counting		Video/Audio Recording; Screen/User View Recording
Words Spoken		Counting		Video/Audio Recording; Screen/User View Recording
Grounding Questions Asked		Counting		Video/Audio Recording; Screen/User View Recording

Mental Workload			Questionnaire (NASA TLX)	
Ease of Use			Questionnaire	
Perception of Outcome and Interaction			Questionnaire	
Pre-Session Expectations			Questionnaire	
Post-session Impressions			Questionnaire	

3.6. Evaluation Setup

The final consideration for the firefighting use case example, once the metrics, measures, and data sources are determined, is how the tasks and evaluation environment will be prepared. To run a usability evaluation of the AR solution with the apartment fire scenario, a testing environment needs to be established with a fire simulated within a space made to look like an apartment. The Test Admin will work closely with the product team to determine details on the testing environment to support conduction of the evaluation and data collection.

Each of the four tasks in the example will require some predetermined information and set up. For example, the "Search location of hazard" task will need multiple rooms in the environment to search. The "Report conditions over radio task" will have pre-determined conditions for the user to communicate such as the location of the fire, or the location of the victim. The "Detect hazards" task will have pre-set hazards for detection such as an open door to a room, combustible material near the fire, and an obstacle blocking a path. The "Monitor air level" task will have predetermined changes in the air level displayed on the HUD.

The physical environment will also need to be properly equipped with video, audio, and streaming devices, data storage and connectivity to collect media data such as user movements, without interfering with users' task performance. The HUD user view will also be recorded, and the headset will log system data needed (i.e., eye tracking, button presses/gestures, and time stamp) throughout the evaluation session.

4. Conclusion

The AR Usability Evaluation Framework provides guidance on planning[2] user-based usability evaluations of AR technology. While this report demonstrates the framework using a firefighting scenario in the public safety domain, the framework is applicable and can be expanded to other domains where user-based AR usability evaluations will be performed.

Applying the framework in AR solution development lifecycles will provide the following benefits:

- Creating explicit structures for user-based evaluations
- Providing a consistent terminology and an initial set of usability metrics
- Facilitating comparability across AR research and development efforts

[2] The AR usability evaluation framework in this report only focuses on planning usability evaluations, and does not cover how to conduct usability evaluations, data analyses and report writing.

- Facilitating sharing of usability evaluation results
- Facilitating establishing human-centered AR design guidelines

As mentioned earlier, a well-conducted and well-planned product development project should follow an iterative human-centered process. Throughout the development lifecycle, iterative, data-driven, user-based usability evaluations with target users should be conducted by trained usability professionals. This report provides a five-component AR Usability Evaluation Framework to facilitate systematic planning of usability evaluations to ensure successful evaluations and collecting useful usability data for product improvement. The five components are: (1) Determine evaluation scope; (2) Identify users and context of use; (3) Develop evaluation scenario and tasks; (4) Select applicable usability metrics; and (5) Define usability measures for selected metrics. Following this framework to conduct usability evaluations throughout development cycle will help reduce development cost and bring AR solutions to market faster while providing usable products that are easy, quick, comfortable, and safe to use.

References

[1] Poetker, B. (2019). A Brief History of Augmented Reality (+Future Trends & Impact). Retrieved March 17, 2022, from https://www.g2.com/articles/history-of-augmented-reality

[2] Javornik, A. (2016). The Mainstreaming of Augmented Reality: A Brief History. Retrieved March 17, 2022, from https://hbr.org/2016/10/the-mainstreaming-of-augmented-reality-a-brief-history

[3] Merriam-Webster. (n.d.). Augmented reality. In *Merriam-Webster.com dictionary*. Retrieved March 17, 2022, from https://www.merriam-webster.com/dictionary/augmented%20reality

[4] ISO 9241-11:2018. *Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts*. ISO, Geneva, Switzerland, 2018, http://www.iso.org/

[5] Choong, Y-Y., and Salvendy, G. (2021). *Voices of First Responders – Applying Human Factors & Ergonomics Knowledge to Improve the Usability of Public Safety Communications Technology: Findings from User-Centered Interviews, Phase 1, Volume 5.* NISTIR 8340, February, 2021. https://doi.org/10.6028/NIST.IR.8340

[6] Theofanos, M.F., Choong, Y-Y., Dawkins, S., Greene, K.K., Stanton, B., and Winpigler, R. (2017). *Usability Handbook for Public Safety Communications – Ensuring Successful Systems for First Responders*. NIST Handbook 161, May, 2017. https://doi.org/10.6028/NIST.HB.161

[7] Hackos, J.T., and Redish, J.C. (1998). *User and Task Analysis for Interface Design*. Wiley Computer Publishing.

[8] Albert, W., and Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. 2nd edition, Morgan Kaufmann.

Appendix A: Apartment Fire Scenario – Complete Task Analysis Flowchart