NIST Internal Report
NIST IR 8489

It's Not Always What the Eye Can See – Challenges in the Evaluation of Augmented Reality

Version 1.0

Kurtis Goad
Kevin Mangold
Susanne Furman

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8489

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8489

It's Not Always What the Eye Can See – Challenges in the Evaluation of Augmented Reality

Version 1.0

Kurtis Goad
Kevin Mangold
Susanne Furman
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8489

September 2023

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

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

NIST IR 8489
September 2023

Publication History
Approved by the NIST Editorial Review Board on 2023-09-13

How to Cite this NIST Technical Series Publication
Goad K, Mangold K, Furman S (2023) It's Not Always What the Eye Can See - Challenges in the Evaluation of Augmented Reality. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Internal Report (IR) NIST IR 8489. https://doi.org/10.6028/NIST.IR.8489

NIST Author ORCID iDs
Kurtis Goad: 0009-0008-0639-2582
Kevin Mangold: 0000-0001-9036-0744
Susanne Furman: 0000-0002-7013-6603

Contact Information
kurtis.goad@nist.gov
kevin.mangold@nist.gov
Susanne.furman@nist.gov

Abstract

Augmented reality (AR) technology is developing at a fast pace. Usability evaluation methodologies for AR need to be updated to accommodate the increasing complexity of AR technology. Eye tracking metrics, which have been historically successful in traditional, often computer-based technology, have been underutilized when evaluating AR devices. Literature review shows documentation of usability methodology used to evaluate AR since 2000. But these reviews did not address the reasons why evaluators have chosen or excluded certain methods, like eye tracking. They also did not address certain evaluations, such as those performed in the industry sector, which are not typically published. This exploratory study addressed these gaps by deploying semi-structured interviews to gain an understanding of the experiences professionals have when evaluating AR technology. Our results show that participants rely most heavily on user feedback and questionnaires/surveys during evaluation. Most participants did not use eye tracking methods. They cited a number of challenges evaluating AR technology and the use of eye tracking, including difficulties in data collection and a lack of consistency and standards across devices.

Keywords

Augmented Reality; Eye Tracking; Usability Evaluation.

Table of Contents

List of Tables

1. Introduction

Augmented reality (AR) technology is evolving faster than the usability methods used to evaluate it. If usability evaluation methods are not researched and updated, we will lose the ability to contribute meaningful analysis of human experiences with new and developing AR technology. This gap would cause AR solutions that come to market to be either unusable or difficult to use for many people. This may erode trust in AR technology and would also slow the pace at which effective AR is developed and made available.

AR has changed drastically since its conception. It began as simple overlays of two-dimensional images onto users' view of their environment, such as a stationary arrow displayed on a piece of glass to point the user in a particular direction. It has since evolved into wearable hardware with software applications that allow people to move freely in physical space. Users can interact with three-dimensional objects that are fully integrated in their physical environment, for example, playing a game where virtual animals walk through their living room.

Research studies have documented the implementation of usability techniques in AR evaluation. In published systematic literature reviews dating from 2005 to 2018 [1–4], results demonstrate that evaluation methodology has not sufficiently evolved to accommodate the increasingly complex use cases of newer AR technology. Existing documented methods frequently overlook the inclusion of multi-task scenarios, rely on a narrow range of metrics, or evaluations are frequently conducted in environments that differ from the intended context of use.

Other research studies [5–19] reveal few documented cases of eye tracking techniques used to evaluate the usability of AR systems in the last 20 years. In evaluations of more traditional computer-based technology (e.g., web sites, computer applications), eye tracking has proven to be an invaluable tool. It offers vital insight without disrupting users' behavior and reduces the need to rely on users to accurately recall and answer questions about what they were looking at after testing [20]. In recent years, some AR devices have included built-in eye tracking technology, and there are commercial eye tracking devices on the market, built specifically for AR devices. Yet, the use of eye tracking metrics used during evaluation remains low. There are numerous documented technical barriers that exist when collecting eye tracking data in AR. These challenges include the complexities of tracking users' gaze at both virtual and physical objects in a unified evaluation [21]; identifying body movements, such as a head turn, to shift gaze [22]; and obtaining precise data in situations where visualizations and physical objects overlap [23]. Even though evidence of these barriers exists, it is unknown if these are the reasons why evaluators do not choose to use eye tracking techniques during evaluation. Understanding the reasons behind evaluators' comparatively lower utilization of eye tracking metrics can provide insights for forthcoming research. As a result, research can then focus on devising solutions to address challenges and introduce enhanced methodologies for evaluating AR technology.

This study identifies a number of gaps that, if addressed, will contribute to the overall improvement of AR usability evaluation methodology. Although there is documentation of methods for usability evaluations and techniques used to develop AR technology, there is not a clear picture of the challenges that practitioners face during evaluation or the reasons behind

their selection or exclusion of techniques. Additionally, a limitation of the documented evaluation methods in current literature is that it is incomplete. Unpublished evaluations that often occur in industry applications are not shared or added to the body of knowledge. Lastly, it is unknown why more evaluators have not chosen to use eye tracking during evaluation.

These gaps are addressed in our research. We conducted a small exploratory study using semi-structured interviews of nine professionals who evaluate AR technology. The following research questions were the focus of our interviews:

RQ1: In what ways have usability professionals been successful in achieving "effective evaluation" of AR technology?

RQ2: Have usability professionals had success using eye tracking metrics while evaluating AR technology? If not, what do they wish to be able to do and how would that help achieve their evaluations goals?

RQ3: What challenges have usability professionals faced that have made it more difficult or prevented them from evaluating AR with the same level of effectiveness as evaluations of more traditional technology?

Our results show that participants most widely rely on formal and informal user feedback and questionnaires/surveys when evaluating AR technology. However, specific methodology choices depend on their research questions and institutional goals. Most participants have not employed eye tracking techniques for evaluation. The reasoning behind their choice to exclude them was not due to the technical barriers cited above [21-23]. The lack of use is because of several challenges including: difficulties collecting and analyzing data, and a lack of consistency and standards across devices. These issues and others were echoed strongly in participants' responses when asked about what challenges they face in performing evaluations as a whole.

2. Methods

We conducted a qualitative interview study of professionals who have worked with AR technology to understand their experiences when performing usability evaluations of AR technology to determine how they could be improved. We developed a semi-structured interview protocol that was reviewed and revised based on subject matter experts' feedback. The National Institute of Standards and Technology Institutional Review Board reviewed and approved the protocol for this project and all subjects provided informed consent in accordance with 15 CFR 27, the Common Rule for the Protection of Human Subjects. We conducted four pilot interviews with participants who were similar to the targeted population to gain feedback related to language, content, flow, and timing of the interview protocol.

2.1. Participants

Participants were federal and non-federal professionals recruited through contacts from NIST employees, and members of the NIST Extended Reality Community of Interest (XR COI) and industry Augmented Reality Enterprise Alliance (AREA)[1]. Participants were required to have

[1] https://thearea.org/

experience evaluating the usability of AR technology, be at least 18 years of age, and could be from any sector (i.e., industry, academia, government). Additionally, participants were not required to have experience using eye tracking techniques during evaluation. Interviews took place between May and June of 2023. The interviews were audio-recorded and professionally transcribed with no personal identifiers linked to the recordings. A total of nine participants were interviewed. Participants included professionals with varying ranges of experience between 2 and 17 years, working with AR from the sectors of industry, academia, government, and military. Participants most commonly had experience working with mobile and head mounted AR systems. They worked primarily with commercial devices, except for one participant whose institution worked with only non-commercial devices. The most common application types that participants had experience with are training, manufacturing, retail, and entertainment. See Table 1 below for a detailed summary.

Table 1. Participant Information

Participant	Sector	System	Device	Applications
P1	Industry	Mobile Head Mounted	iPad[2] iPhone Android HoloLens Google Glass	AR Web Browser Games Retail Manufacturing/Service Documentation
P2	Industry	Mobile	Smartphone	3D Mapping Entertainment Retail
P3	Academia	Mobile Head Mounted Audio	Smartphone Google Glass HoloLens HoloLens 2 Magic Leap Meta Quest Pro	Office Applications Education Training
P4	Academia	Mobile Head Mounted	Smartphone HoloLens Magic Leap Magic Leap 2 Meta Quest Pro	Retail Navigation Education Entertainment Training
P5	Industry	Mobile Head Mounted	Tablet HoloLens Magic Leap Google Glass	Manufacturing Communication
P6	Academia/Government[3]	Mobile Head Mounted	HoloLens HoloLens2 Google Glass Nreal Light Magic Leap Magic Leap 2 Vuzix Varjo XR-3	Scientific Visualization Training 3D Modeling Digital Engineering Disaster Assessment Collaboration

[2] Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.
[3] Participant is primarily employed in academia, but their work is sometimes involved with the government sector.

			Campfire Meta Quest Pro	
P7	Industry	Mobile Head Mounted Computer	iPad iPhone HoloLens 2	Manufacturing Training
P8	Military	Head Mounted Head Up Computer	Non-Commercial	Military
P9	Government/ Academia[4]	Mobile Head Mounted	Android HoloLens HoloLens2 Magic Leap Magic Leap 2 Meta Quest Pro Varjo XR-3 Vuzix Blade	Public Safety Image Detection Test Beds Spatial/Location mapping Training

2.2. Data Analysis

We used both deductive and inductive coding of the interview transcripts. We developed an a priori code list based on research questions, literature review, and meetings with content area experts. Two team members used the initial code list to code the same three interviews, allowing for identification of points of agreement and disagreement. Discussion around areas of disagreement provided a greater ability to refine codes and pursue alternative interpretations of the data. The team identified emergent codes via the first round of coding and added to the code list for use with subsequent interviews. Research team discussions occurred frequently to review team members' use of codes, identification of additional emergent codes, and issues found with the use of the code list. Revisions were made based on these discussions. The remaining interviews were coded independently by team members. Once all interviews were coded, research team members met to identify points of crossover, patterns in the coded data, and abstract themes that grew out of the coding.

2.3. Limitations

This study focused on examining the experiences of participants who evaluate the usability of AR technology. The nature of the study was exploratory, focusing on a small sample of qualified participants, in order to gather data to inform future research. For this reason, the results of this study cannot be generalized to the population of interest as a whole. Rather, results will inform the direction of future projects to gain a more detailed and comprehensive understanding of the experiences of people who evaluate AR technology.

Recruiting representative participants was a challenge, and this may be a bigger issue with future research aimed at recruiting a larger sample. The participants chosen for the study were selected from what could be considered a narrow population. AR technology is not as widely adopted as

[4] Participant is primarily employed in government but also does related work in academia.

other technologies such as computers, non-AR smartphone applications, and others. For that reason, it can be inferred that not as many people are evaluating the usability of AR. Additionally, since recruitment materials specified professionals with or without experience using eye tracking could participate, it is possible that some potential participants who fit the study criteria did not participate due to a lack of experience with eye tracking techniques.

3. Results

In this section, we present interview findings. Participant quotes are provided with participant identifiers and timestamps (e.g., (P1 – 11:54) indicates a quote from Participant 1 at 11 minutes and 54 seconds). Counts are provided in the presentation of the interview data. These counts are not meant to be expressed as quantitative data that can be generalized to the population studied. Instead, they are meant to demonstrate the weight of certain responses and how some sentiments were shared or not shared across participants.

3.1. Evaluation Methods

In the following section, we discuss participant responses related to test environments, frequently chosen methods for evaluating AR, and any experiences using eye tracking evaluation methods.

3.1.1. Test Environment

When asked about their choice of test environments, all participants stated that they initiate evaluations in a lab or lab-like environment. Some participants indicated that they choose to also perform evaluations in the intended environment of use. This decision is influenced by several factors. Participants within the industry sector, when assessing manufacturing applications, emphasized the significance of obtaining user feedback within their manufacturing environments. This practice ensures that their investment in AR technology aligns with advancing the achievement of their company's objectives.

> "…we kind of take any of these applications and things like that, and we've seen the demos and we've seen them run perfectly, but we very commonly need to bring it back, bring it into our environments, and see how well it works." (P7 – 06:20)

> "We're not really in the business of building AR. Excuse me. We're in the business of building products. So for us, the whole reason we're investing in this technology is for those reasons, is for building more product, faster, to a higher degree of quality. So that's really how we measure why we're doing what we're doing." (P5 – 11:24)

Many participants cited limitations of the AR technology they evaluate as a barrier to being able to test in some environments.

> "It's primarily been lab studies and for the main reason of how the devices tend not to work as well outside. High glare situations." (P3 – 10:15)

For others, it depends on the research question being asked.

> "I'm conducting an evaluation with Google Maps AR. So of course, that has to be done outside. So I've done that. There were a couple of retail app that was [*sic*] also conducted outside because it was looking at outdoor products that someone can place down." (P4 – 07:45)

Testing in the environment of use can also be more difficult for some use cases of AR. For example, a participant who evaluates applications intended to be used by the military in combat situations states:

> "It would be nice to do more fielded stuff. But it gets tough to control things. You just need a lot of that control to get feedback on kind of precise overlay features or details, whatever it is, that you lose control of in the field." (P8 – 19:53)

Another participant, who works with entertainment applications, does not share the same difficulties when moving their evaluations from the laboratory to the intended environment of use:

> "But [Company Name][5], specifically, they're creating toys and games, and of course, we're testing these in an environment. And the beauty of AR is you can really test it anywhere, so we always test it just outside. We just go outside, we put a QR code on the wall, set up a domain, and then we test it." (P2 – 16:30)

3.1.2. Methods and Metrics

When participants were asked about what methods they most frequently choose, the majority of participants stated user feedback, and questionnaires/surveys. Participants reported collecting user feedback from informal user comments, think aloud procedures, and interviews. Many feel that getting feedback directly from users is the quickest and clearest way to understand improvements that need to be made.

> "The methodology we use is that we create alpha or beta versions. We put them in the hands of customers. We take back feedback from the customer, and there are certain requirements that we're looking to satisfy." (P1 – 07:48)

> "So I feel like we get a lot of rich information from users' comments as they're using the device and into what goes well, what's surprising, what

[5] The company name mentioned here was redacted to protect the identity of the participant.

> is confusing more so than you can get out of a questionnaire or even a focus group" (P4 – 16:50)

> "It's a lot of making sure who's going to use it is going to want to use it because as soon as we make anybody's job harder, they're just not going to adopt the technology, even if it could save tons of time or headaches down the road." (P7 – 15:40)

Participants that choose questionnaires/surveys are seeking user perceptions of ease of use, comfort, and satisfaction, as well as data not easily obtained from direct user feedback, (for example, mental workload and presence). These participants feel that questionnaire data are easier to obtain and more dependable in certain situations (for example, for publishing their data and discovering insights that may contradict user comments).

> "I think most common is questionnaires to a certain extent. I think they're just, I want to say, tried and true. But they're more readily deployed. I think the researchers are more comfortable in leveraging them. And in some cases, I'll say that they've been validated. And so I feel like the researchers are able to take that approach and have higher confidence in their results or at least in trying to create a publication from it." (P9 – 13:49)

> "We found a lot of the time, people say, "I'm fine, and I feel great after using this AR device." But the simulation sickness questionnaire says otherwise, that they have a lot of ocular symptoms, which is really surprising, which is why we tend to collect that data a lot." (P4 -16:50)

One participant, responsible for evaluating AR applications in manufacturing, prioritizes time on task and error counts above all other metrics. This is because they focus on assessing the effectiveness and efficiency of the AR technology as an aid for performing manufacturing tasks.

> "Oh, by far, it would be the time to complete jobs and the number of defects recorded. So I didn't really talk about the defects, but obviously, quality inspectors will come by off every job, and they'll either accept or reject it. And that will get recorded in the CMES [Common Manufacturing Execution System] system. So we also track the number of quality pickups when we're using AR versus not." (P5 – 10:44)

3.1.3. Eye Tracking

Direct experience using eye tracking as a method for evaluating AR was limited to one participant. They used a device's onboard eye tracker to capture eye tracking data and reviewed a playback of the user's view as well as fixation and dwell time data captured by the device.

> "In a lot of cases, yeah, we're just playing it back and seeing what they're interested in, where their eye is fixated. As the research gets more

> formalized, that's when we start putting kind of tags on it, right? What are they looking at? And so in those cases, we've been using the onboard eye tracking so that the game engine knows specifically what it is that they are looking at, and so we include that in the data dump to disk. And so it's very easy to count and figure out how long they've been fixated at certain objects." (P6 – 23:34)

One participant cited a collaborator's use of pupil dilation data as an indicator of cognitive load.

> "From collaborators, they've used it as that real-time cognitive load indicator, specifically the pupil dilation. And they've had pretty good success with that." (P9 – 13:49)

Eight out of nine participants had not used eye tracking during evaluation. By far the most expressed reason for not using it was that it is difficult to use. Collecting eye tracking data takes more time and effort than other data. Without a clear advantage to what they will gain from the data, participants have chosen not to collect it.

> "I know it's possible, but until that's very easy at the UX level, just to grab that along with the performance data you may be getting, or some subjective ratings that you may be getting, then it's a little bit of a barrier." (P3 – 22:46)

> "I don't know if we have a good description of or what benefit we would get out the other end…then it just becomes a weighing out of do we want to go get that info, what benefit are we going to get versus the time it would take to implement those features." (P7 – 23:42)

It is also difficult to implement eye tracking depending on the use case of the AR application being evaluated. For example, in a manufacturing use case where users are more physically engaged in their task while using a mobile AR device not attached to their head, collecting eye tracking data could prove to be more cumbersome than beneficial.

> "So the person was walking around a whole room in this study. They were collecting parts, they were going to…collect tools, they were doing the assembly, looking at [a] tablet. And they weren't wearing a headset. They were using a tablet for instruction. So I guess just ease." (P5 – 24:02)

Another issue arises from the lack of consistency between AR technologies. There are many different AR devices, and there are differences in how those devices implement features and allow access to data. This creates a challenge for evaluators when trying to create consistent methodology for using eye tracking data across devices.

> "But as I understand it, [it's] really hard to pull out some of that raw data. They don't give you access potentially because of privacy concerns, sensitivity concerns, either on the end user or on the company as a whole on how they're developed and deploy the technology. But I think it's that

> lack of consistency that can be kind of challenging to deploy the solutions." (P9 – 27:46)

When asked how participants wish they could use eye tracking during evaluation, numerous metrics commonly associated with eye tracking were mentioned such as fixations, dwell time, gaze path, and cognitive load. However, the prevailing sentiment among many participants was the desire for easier access to data and simplified analysis.

> "Ideal world, yeah. So it works. There's a very short calibration process, if any. It's a continuous calibration process. The data is not noisy. It's easy to pull out that raw data." (P9 – 30:13)

> "So I'd certainly like to see more or easier access to pupillometry data. Some of the eye trackers have better access to that than others." (P6 – 25:59)

> "I mean, the toughest part of eye tracking in my opinion is the analysis, right? So what I would want is given my scene, make it super easy to establish where my areas of interest are. I run my participant and [it] just spits out what my data of interest [are]." (P3 – 24:31)

> "If there is a way with AI or something that can help analyze that data a bit more seamlessly or just a way to capture movements…It just becomes very, very tedious. So I cannot think of a way, but [it would] be great if there was a way to make that more seamless, easier to do." (P4 – 29:53)

Participants' concerns regarding the use of eye tracking were just part of a much broader discussion about the challenges practitioners encounter when evaluating AR technology.

3.2. Challenges

We asked participants to describe challenges or difficulties when evaluating AR. In this section we present the findings from the two most shared challenges: collecting and analyzing data, and a lack of consistency and standards. Then, we present additional challenges, which, while not as commonly shared, provide a context for exploration in future research.

3.2.1. Collecting and Analyzing Data

Across participants, the most commonly shared challenge was collecting and analyzing data. For some methods, like eye tracking discussed above, the data is unique. Therefore, it is more difficult to seamlessly incorporate into data collection and analysis. One participant uses the example of physiological data:

> "I wish physiological data was easier to deploy, it wasn't as noisy, and was more consistent across all the devices." (P9 – 19:49)

In other instances, the use case for the AR application participants evaluate may be more complicated and involve complex technology. In this case, participants feel that more data is required for accurate evaluation.

> "The biggest data collection we did was over about two or three thousand hours-worth of work, mechanics labor, and even then I felt that we could have used more." (P5 – 20:40)

Some participants simply want more quantitative data during evaluations. They feel it is easier to support their conclusions with quantitative data but that it is harder to obtain with AR applications.

> "I do wish a lot more of it was quantitative. I don't know if I can say how it would be. It would be way easier, especially with most of the people that we work with, that kind of engineering mindset that if there was a number to it, it's easier to pitch, it's easier to prove against other softwares [*sic*], other hardware, anything like that." (P7 – 19:46)

> "We had greater ambitions for the users to kind of complete a series of tasks within the application and to extract that data and have that as kind of the quantitative back end. Ultimately, we had to defer to just capturing, "What is the iteration? What is the instance of the solution that we're going to give them?" (P9 – 18:50)

An issue associated with data collection, mentioned by participants, is the presence of institutional restraints. Some companies may not prioritize spending resources to collect certain types of data, even if it may benefit the evaluation process.

> "We don't usually, for instance, gather telemetry on where did they move in the physical space when they were using the software or auditing every single button click and things like that. We just don't tend-- I haven't seen those kind of instances where we frankly have the resources to do that." (P1 – 34:12)

For some, it may be possible to collect the data, but the process required by their institution to do so tends to delay the evaluation process or act as a deterrent, discouraging regular practice of certain methods.

> "Because then we're doing human subject research, right, and we are measuring how somebody goes through it. Now, the IRBs in this case, usually when we do that, the hardware we're using it as per the manufacturer's spec, it's all within the norms, and so we can get-- there is an expedited process…But it does add oversight and overhead for research projects." (P6 – 10:40)

> "But we've run into a number of issues with just data storage, and our own personal constraints being government and military and using a lot of those technologies that kind of speed up that process." (P8 – 10:14)

Another challenge, mentioned by participants across each sector, was the time required to perform formal evaluations. This challenge overlaps with difficulties collecting and analyzing data because it is often the time required to collect certain data that causes issues for participants.

> "So quantitative studies take a lot longer to collect data that way, for us, just the way to procure soldiers is not as seamless as we would like it to be." (P8 – 12:20)

> "And so if we start gathering that kind of data, it does have an impact on our timelines." (P6 – 10:40)

> "…timing is very important, and you only have the participant for a certain amount of time. So if you're fighting with the technology, then that wastes time." (P3 – 17:50)

Time spent on evaluations can also be a sensitive business decision when stakeholders are involved. As one participant stated:

> "We will recruit our customers to help us. They're often eager to help us. But there's limits to that. I mean, if it's going to require a lot of time on their part, that costs them money. So we're sensitive to that. Even when I was at a startup, the idea of asking a company to dedicate hours and hours of their employees' time to help us develop our products was never a winning strategy." (P1 – 35:31)

3.2.2. Lack of Consistency and Standards

The second most frequently mentioned challenge among participants was the lack of consistency and standards across AR devices. It is not missed by participants that, at a fundamental level, any given application is limited in its usability if it is restricted to a single device.

> "It's just the wild, wild west out there…Because from a UX standpoint, we shouldn't be looking at just a HoloLens or just the Magic Leap. It's really here's my application. I don't care what device I'm going to run it on. It should work everywhere. And then I can get a true sense of the true UX of this application." (P3 – 28:12)

This is a pain point for evaluators, especially when compared to the ease of evaluating web pages, which are accessible across different devices and seem to have clearer requirements.

> "…augmented reality is hardly-- even as normalized as something like having a webpage for your company. I mean, in the early days of having

> web pages, at least, pretty quickly, we realized what a company might need. They need to have a face out there. They need to answer certain basic questions. And then augmented reality is all over the map." (P1 – 36:53)

In addition to being a fundamental underlying issue, the lack of consistency and standards also creates additional challenges at various evaluation points. Examples given by participants include creating documentation for applications, comparisons between device types, and device accuracy.

> "One of my important beliefs about this field is that there needs to be a certain consistency between what we call interactive documentation… All of those really should come from the same source. You should be authoring all that at the same time if you really want to gain some efficiencies." (P1 – 22:24)

> "I do wish that there were some good standards in how some of the display resolution and density was provided….Man, I would love…a great standards document that says, 'Here are the different display technologies in use, and here are various metrics that you want to use to compare these classes of devices,' because comparing a-- I'm not sure that a Quest Pro augmented reality field of view measurement and a HoloLens 2 field of view measurement is necessarily an apples-to-apples comparison." (P6 – 16:15)

> "Because not to belabor this too much, but in certain circumstances, let's just take the HoloLens. The accuracy varies wildly. If you're in a very blank space with not much to track, you're going to get very poor accuracy. If you're in a dense space-- so we had to really understand in our environment, how accurate could it be?" (P5 – 18:58)

Participants also have difficulties—due to their own institutional policies as well as those of the companies that develop AR devices—that are exacerbated by the lack of consistency and standards across devices. Participants' institutions may have security policies that mandate AR devices to meet specific standards, and certain devices might align better with these polices than others. Additionally, the companies developing AR devices have diverse standards concerning how data can be accessed from their devices, creating challenges for evaluators who seek to utilize that data.

> "And then more specifically, for our company and the defense industry, as soon as you start to get into the applications and even the hardware as well, there's a lot of stuff with-- they have to meet our security concerns…and then we run into issues with what type of OS [operating system] are the things running, and there's just some that are a lot harder to integrate than others, so. I don't know. It's a challenge from start to finish, but I think we're getting there." (P7 – 17:33)

> "But as I understand it, [it's] really hard to pull out some of that raw data. They don't give you access potentially because of privacy concerns, sensitivity concerns, either on the end user or on the company as a whole on how they're developed and deploy the technology. But I think it's that lack of consistency that can be kind of challenging to deploy the solutions." (P9 – 27:46)

3.2.3. Additional Challenges

Other challenges, while not widely shared among participants, still hold significant importance as they might emerge more prominently in further and more extensive research. Some of these challenges are associated with known physical issues with AR hardware, such as motion sickness, and comfort.

> "If you start getting a headache, who gets a headache, how bad is the headache, and how many people do you need to put in that headset before you start having a good enough sample size to know whether or not that experience is causing that headache, or if…a barometer change that day is giving somebody a headache?" (P6 – 13:29)

Another challenge is accounting for users' lack of familiarity with AR devices and the impact on results.

> "If someone hasn't really used the device before or used it once or twice, you're not really getting much information from that." (P4 – 16:50)

A lack of experience specifically evaluating AR technology can pose a challenge.

> "We do have usability people at [Company Name], but if you tell them you're working with the HoloLens, they're kind of like, 'Well, okay, that's out of my—' so I don't know." (P5 – 15:17)

Some AR applications tend to be more complex than screen-based technology, they are more prone to create technical issues for evaluators that extend the pilot study process.

> "So it's a matter of maybe piloting a lot more than what we might be doing if it were not AR." (P3 – 17:50)

The wide variety of use cases for AR applications necessitates more critical considerations when determining the data that needs to be collected and utilized to assess the success of the AR system.

> "In some areas that are very mission critical, whether it's nuclear engineering or something like that, if you were going to-- if you were going to push an augmented reality application to those types of people, I think you might have to really-- you would really have to run through some things with actual users identifying potential mistakes and errors that can arise or usability issues, confusion, etc., whereas if you're

> cleaning up garbage or sorting recyclables, it may not be that important...” (P1 – 38:09)

Another challenge, mentioned primarily by participants in the industry sector, is the accuracy of device performance.

> “So we work indoors, we're centimeter accurate, and so this is a very tough nut to crack.” (P2 – 20:10)

> “Any time anyone new sees the HoloLens or whatever, the first question is well, how accurate is it? So if you're going to show someone where to drill a hole or where to write a wire or do something, you have to have some quantifiable answer.” (P5 – 17:34)

> “I think it's difficult because-- I mean, a big part of it is nothing that we've seen places itself accurate enough to get great measurements off of.” (P7 – 09:08)

As previously described, participants in the industry sector confirmed that they conduct evaluations, even if only informal ones, in the intended environment of use. Data from our interviews revealed that challenges related to evaluating in the intended environment of use were only raised by non-industry participants.

> “It's primarily been lab studies and for the main reason of how the devices tend not to work as well outside. High glare situations.” (P3 – 10:15)

> “I guess I'm sure you know with the Meta Quest Pro, if the sun hits it, it's completely destroyed. So those are some issues that are out there that just the hardware providers need to work towards and work on making it a bit better.” (P4 – 34:26)

> “I mean, obviously…[for] the generalizability of lab studies, it would be nice to do more fielded stuff.” (P8 – 19:53)

> “The most recent one that we did was done on scene. However, the scene was in a theater room while the event was unfolding around the user. And in this case, we evaluated those solutions under fairly ideal conditions. I'd say it wasn't outside. It wasn't in direct sunlight. You weren't worried about UV or having to outshine the sun and worry about contrast and other effects there with these devices.” (P9 – 09:39)

4. Discussion

RQ1: In what ways have usability professionals been successful in achieving “effective evaluation” of AR technology?

Participants widely prioritize collecting user feedback and utilizing questionnaires and surveys during evaluation. This reliance on swift and straightforward methodologies is a logical answer to the myriad of challenges evaluators face while evaluating AR technology – such as difficult data collection and analysis, a dizzying lack of standards across devices, and the need to negotiate time spent on evaluation. User feedback and questionnaires rely solely on user interactions and avoid the technical obstacles that other methods, contingent on the performance of the AR system, can encounter. For example, it is easier and less time/resource consuming to ask a user how well an AR device helped them perform a task than it is to successfully link an AR device to a computer so performance data can be gathered without error.

The choice of methods also depends on the research questions or evaluation goals, which vary by the sector. For example, one participant in the academic sector employs a systematic approach to selecting methods based on the research question and obtaining as much data as possible to support their evaluation outcomes:

> "I mean, I think initially when we have a research question or somebody comes to us to say, hey, can you help us evaluate this? We try and figure out what do we need, right? And then based on what do we need, how are we going to get there? What are the methods that we can use? And then…knowing the weaknesses of some of these, the technology, then we'll design accordingly. Or most of our studies, we triangulate, right? So we get as many objective measures at the same time subjective so that we have a lot of data to sift through and not just rely on one measure and say, this is our answer." (P3)

Another participant in the industry sector prioritizes methods that zero in on measuring the success of the AR technology in increasing the quality of work for specific tasks.

> "Oh, by far, it would be the time to complete jobs and the number of defects recorded." (P5 – 10:44)

These domain differences may help to explain the participants' choices of test environments. Participants in the industry sector often seek informal feedback and use only a few specifically directed measurements of user performance; they may find it easier to evaluate in the intended environment of use. For example, participants that evaluate manufacturing applications are typically seeking feedback about devices that are already being deployed on the factory floor. Other participants, whose evaluations tend to be more research oriented and include various AR applications, may rely more on laboratory settings where they can use evaluation strategies for collecting large amounts of data.

Future research, aimed at collecting responses from a larger sample of evaluators, should also focus on obtaining participants from each sector in order to further explore difference in evaluation strategies.

RQ2: Have usability professionals had success using eye tracking metrics while evaluating AR technology? If not, what do they wish to be able to do and how would that help achieve their evaluations goals?

The difficulty in determining the optimal path for improving AR evaluation methodology lies in past research efforts. These research efforts only document the methods chosen or rejected and seldom delve into the underlying reasons behind practitioners' decision-making. This leads to the question of why a greater effort has not been made to use historically successful methods, such as eye tracking. As one participant said:

> "I can see where you're coming from. I mean, you're probably well aware that eye tracking has become an indispensable tool for evaluating screen-based software …Naturally, the same technology should be valuable in evaluating AR, so I can appreciate that angle." (P1 – 56:09)

Our conversations with participants revealed that there are more pressing challenges that must be addressed before having the luxury of incorporating new methods into their repertoires. As discussed earlier, previous research highlights technical barriers to using eye tracking effectively as an evaluation method for AR [21-23]. However, none of the participants mentioned encountering those technical barriers. The reasons for participants' avoidance of using eye tracking were related to challenges they experienced before any of the technical barriers became an issue. Even if these documented technical barriers of eye tracking technology were resolved, it is evident that participants may still be hesitant to use it during evaluation.

The reasons participants stated for not using eye tracking evaluation methods were related to the difficulty of data collection/analysis and the lack of consistency between devices. Issues with eye tracking are just part of the challenges in the evaluation of AR technology as a whole. Future research should explore these challenges, outside of the context of eye tracking. The goal is to identify the most crucial challenges, which if addressed, will significantly enhance the efficacy of evaluating AR technology.

RQ3: What challenges have usability professionals faced that have made it more difficult or prevented them from evaluating AR with the same level of effectiveness as evaluations of more traditional technology?

The two most widely shared challenges were difficulties collecting/analyzing data and a lack of consistency and standards for AR devices. These challenges are often intertwined in participant responses. Unique data, like eye tracking and physiological data, are inherently more difficult to collect on AR devices because they are often collected using technology sensitive to movement. The strategies for collecting this type of data tend to differ due to disparities among AR devices, making the data less "consistent across all devices" (P9 – 19:49).

Participants also cited institutional limitations that create challenges related to both data collection and the lack of consistency across devices. Institutional policies may delay and lengthen the process of collecting certain data and prevent participants from "using a lot of those technologies that kind of speed up that process" (P8 – 10:14). Difficulties also exist for participants where institutional policies intersect with policies of AR device manufacturers, both of which require certain privacy and security standards be met. Because the policies of AR device manufacturers can vary, it is often difficult for participants to use a variety of devices that consistently meet their own institutional policies for privacy and security.

The need for more dependable data is expressed in participants desire to collect more quantitative data—data that although harder to obtain, better supports their conclusions. The

same is true for participants who want more data in general. Use cases for AR applications often encompass a variety of intricate tasks, such as those found in manufacturing. As a result, participants feel that the data they collect does not sufficiently represent all potential applications of the devices they are evaluating. However, there are certain pain points, created by the lack of consistency and standards across devices, that play a role in limiting participants ability to expand their data collection. These pain points include a lack of consistent use of instructional and informational visualizations (interactive documentation), difficulties comparing different AR device types, and variable device accuracy.

5. Conclusion

In a small exploratory study, we interviewed nine professionals who evaluate the usability of AR technology to understand their evaluation experiences and gain insight on how the process of evaluating AR can be improved. Results indicated that participants widely rely on user feedback and questionnaires/surveys during evaluation. There were few examples of participants expanding their methods to include other methods, such as eye tracking. This is due to a number of challenges evaluators encounter, primarily relating to difficulties collecting and analyzing data, and a lack of consistency and standards across devices.

Given the constraint of a limited participant pool in this study, it is not possible to assert the significance of these challenges to the entire community of AR evaluators. Nonetheless, this data proves valuable in guiding future research endeavors that seek to delineate the challenges faced by the AR community for the purpose of improving AR technology evaluation methodology.

References

[1] Swan JE, Gabbard JL (2005) Survey of User - Based Experimentation in Augmented Reality. *Proceedings of 1st International Conference on Virtual Reality*:1–9. https://doi.org/10.1.1.527.6345

[2] Dünser A, Grasset R, Billinghurst M (2008) Survey of Evaluation Techniques Used in Augmented Studies. *ACM SIGGRAPH ASIA 2008 Courses, SIGGRAPH Asia'08* (June 2014). https://doi.org/10.1145/1508044.1508049

[3] Bai Z, Blackwell AF (2012) Analytic review of usability evaluation in ISMAR. *Interacting with Computers* 24(6):450–460. https://doi.org/10.1016/j.intcom.2012.07.004

[4] Dey A, Billinghurst M, Lindeman RW, Swan JE (2018) A systematic review of 10 Years of Augmented Reality usability studies: 2005 to 2014. *Frontiers Robotics AI* 5(APR). https://doi.org/10.3389/frobt.2018.00037

[5] Chen MC, Lim V (2013) *Tracking Eyes in Service Prototyping* eds Kotze P, Marsden G, Lindgaard G, Wesson J, Winckler M (Madeira Interact Technol Inst, Funchal, Portugal NR - 9 PU - SPRINGER-VERLAG BERLIN PI - BERLIN PA - HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY), Vol. 8120.

[6] Dzsotjan D, Ludwig-Petsch K, Mukhametov S, Ishimaru S, Kuechemann S, Kuhn J, MACHINERY AC (2021) *The Predictive Power of Eye-Tracking Data in an Interactive AR Learning Environment* (TU Kaiserslautern, Dept Phys, Kaiserslautern, Germany). https://doi.org/10.1145/3460418.3479358 WE - Conference Proceedings Citation Index - Science (CPCI-S)

[7] Josephson S, Myers M (2019) Augmented Reality Through the Lens of Eye Tracking. *VISUAL COMMUNICATION QUARTERLY* 26(4):208–222. https://doi.org/10.1080/15551393.2019.1679636

[8] Kim KH, Wohn KY (2011) Effects on productivity and safety of map and augmented reality navigation paradigms. *IEICE Transactions on Information and Systems* E94-D(5):1051–1061. https://doi.org/10.1587/transinf.E94.D.1051

[9] Kim SJ, Dey AK (2016) Augmenting human senses to improve the user experience in cars: applying augmented reality and haptics approaches to reduce cognitive distances. *Multimedia Tools and Applications* 75(16):9587–9607. https://doi.org/10.1007/s11042-015-2712-4

[10] Kluge M, Asche H (2012) *Validating a Smartphone-Based Pedestrian Navigation System Prototype An Informal Eye-Tracking Pilot Test* eds Murgante B, Gervasi O, Misra S, Nedjah N, Rocha A, Taniar D, Apduhan BO (Univ Potsdam, Dept Geog, Karl Liebknecht Str 24-25, D-14476 Potsdam, Germany NR - 20 PU - SPRINGER-VERLAG BERLIN PI - BERLIN PA - HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY), Vol. 7334.

[11] Park KS, Cho IH, Hong GB, Nam TJ, Park JY, Cho SI, Joo IH (2007) *Disposition of information entities and adequate level of information presentation in an in-car augmented reality navigation system* eds Smith MJ, Salvendy G (Korea Adv Inst Sci & Technol, Dept Ind Engn, Taejon, South Korea), Vol. 4558.

[12] Renner P, Pfeiffer T, Marchal M, Teather RJ, Thomas B (2017) *Attention Guiding Techniques using Peripheral Vision and Eye Tracking for Feedback in Augmented-Reality-Based Assistance Systems* (Bielefeld Univ, Cluster Exellence Cognit Interact Technol, Inspirat 1, D-33619 Bielefeld, Germany FU - Cluster of Excellence Cognitive Interaction Technology “CITEC” at Bielefeld University [EXC 277]; German Research Foundation (DFG)German Research Found).

[13] Ramkumar N, Fereydooni N, Shaer O, Kun AL (2019) *Visual Behavior During Engagement with Tangible and Virtual Representations of Archaeological Artifacts* eds Cauchard J, Gentile V, Khamis M, Sorce S (Univ New Hampshire, Durham, NH 03824 USA). https://doi.org/10.1145/3321335.3324930

[14] Pfeiffer J, Pfeiffer T, Greif-Winzrieth A, Meissner M, Renner P, Weinhardt C (2017) *Adapting Human-Computer-Interaction of Attentive Smart Glasses to the Trade-Off Conflict in Purchase Decisions: An Experiment in a Virtual Supermarket* eds Schmorrow DD, Fidopiastis CM (Karlsruhe Inst Technol, Karlsruhe, Germany), Vol. 10284. https://doi.org/10.1007/978-3-319-58628-1_18

[15] Rohs M, Schleicher R, Schoning J, Essl G, Naumann A, Kruger A (2009) Impact of item density on the utility of visual context in magic lens interactions. *PERSONAL AND UBIQUITOUS COMPUTING* 13(8):633–646. https://doi.org/10.1007/s00779-009-0247-2

[16] Sayed AM, Shousha MA, Islam MDB, Eleiwa TK, Kashem R, Abdel-Mottaleb M, Ozcan E, Tolba M, Cook JC, Parrish RK (2020) Mobility improvement of patients with peripheral visual field losses using novel see-through digital spectacles. *PLOS ONE* 15(10). https://doi.org/10.1371/journal.pone.0240509

[17] Tanabe A, Yoshioka Y (2020) *Gazing Pattern While Using AR Route-Navigation on Smartphone* ed Ahram T (Chiba Univ, Grad Sch Engn, Inage Ku, 1-33 Yayoi Cho, Chiba, Chiba, Japan FU - JSPS KAKENHIMinistry of Education, Culture, Sports, Science and

Technology, Japan (MEXT)Japan Society for the Promotion of ScienceGrants-in-Aid for Scientific Research (KAKENHI), Vol. 973. https://doi.org/10.1007/978-3-030-20476-1_33

[18] Wang TK, Huang J, Liao PC, Piao Y (2018) Does Augmented Reality Effectively Foster Visual Learning Process in Construction? An Eye-Tracking Study in Steel Installation. *Advances in Civil Engineering* 2018. https://doi.org/10.1155/2018/2472167

[19] Wiesner CA, Ruf M, Sirim D, Klinker G (2017) *3D-FRC: Depiction of the future road course in the Head-Up-Display* eds Broll W, Regenbrecht H, Swan JE (Robert Bosch GmbH, Stuttgart, Germany). https://doi.org/10.1109/ISMAR.2017.30

[20] Tullis T, Albert B (2013) Chapter 7 - Behavioral and Physiological Metrics. *Interactive Technologies*, eds Tullis T, Albert BBT-M the UE (Second E (Morgan Kaufmann, Boston), pp 163–186. https://doi.org/https://doi.org/10.1016/B978-0-12-415781-1.00007-8

[21] Gardony AL, Lindeman RW, Brunye TT (2020) *Eye-tracking for human-centered mixed reality: promises and challenges* eds Kress BC, Peroz C (US Army Combat Capabil Dev Command Soldier Ctr, 10 Gen Greene Ave, Natick, MA 01760 USA), Vol. 11310. https://doi.org/10.1117/12.2542699

[22] Alao N (2020) *Qualitative and Quantitative Visual Information Detected by Portable Eye-Tracking Technology* eds Kress BC, Peroz C (FAUL Lisbon Sch Architecture, Dept Drawing Geometry & Computat, Lisbon, Portugal NR - 11 PU - SPIE-INT SOC OPTICAL ENGINEERING PI - BELLINGHAM PA - 1000 20TH ST, PO BOX 10, BELLINGHAM, WA 98227-0010 USA), Vol. 11310. https://doi.org/10.1117/12.2548336 WE - Conference Proceedings Citation Index - Science (CPCI-S)

[23] Elmadjian C, Shukla P, Tula AD, Morimoto CH (2018) *3D gaze estimation in the scene volume with a head-mounted eye tracker* ed Spencer SN (Univ Sao Paulo, Comp Sci Dept, Sao Paulo, Brazil). https://doi.org/10.1145/3206343.3206351 WE - Conference Proceedings Citation Index - Science (CPCI-S)