11-12-2021

The Challenges of Evaluating the Usability of Augmented Reality (AR)

Jessyca L. Derby
Embry-Riddle Aeronautical University, derbyj1@my.erau.edu

Barbara S. Chaparro
Embry-Riddle Aeronautical University, chaparb1@erau.edu

Scholarly Commons Citation

Derby, J. L., & Chaparro, B. S. (2021). The Challenges of Evaluating the Usability of Augmented Reality (AR). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 65*(1). 10.1177/1071181321651315

The Challenges of Evaluating the Usability of Augmented Reality (AR)

Jessyca L. Derby, & Barbara S. Chaparro
Department of Human Factors and Behavioral Neurobiology
Embry-Riddle Aeronautical University, Daytona Beach, FL

Augmented reality (AR) is a new and emerging technology that could benefit from evaluating its usability to better the user's experience with the device or application. This is often done through usability testing and heuristic evaluations. However, AR technology presents some challenges when completing these usability evaluations. Practitioners need to keep in mind the hardware limitations of AR devices that may not be present with other computerized technology, consistency of the users' environment plays a larger role in the AR experience, recognize that a novelty effect may occur and affect subjective scores, and choose heuristic sets that will best evaluate AR applications. Practitioners need to be aware of these challenges and overcome them to accurately assess the usability of these products to gain insights about what should be changed to make the overall experience with the product better.

INTRODUCTION

Augmented reality (AR) has become more entwined in our everyday lives since its introduction in the late 1960s (Sutherland, 1968). It has been used as a training tool, can give us directions through Google Maps, is used as product marketing, included in games such as Pokémon Go, and social media platforms like Snapchat (Barsom et al., 2016; Vilkina & Klimovets, 2019; Snap Inc, 2019; Google LLC, 2020; Niantic, 2020). AR can be defined using Milgram's Reality-Virtuality Continuum, pictured in Figure 1 (Milgram et al., 1995). One side of the continuum is reality as we see it, and on the other side is a fully virtual environment, which includes technology such as Virtual Reality (VR) where the user is fully immersed in the virtual environment. Augmented reality (AR) and augmented virtuality (AV) are in between the two ends of the spectrum as they blend the real and virtual worlds together. AV incorporates aspects of the real environment into a virtual environment, such as displaying the location of the user's hands in a VR simulation. AR differs, as it adds virtual information onto their physical environment.

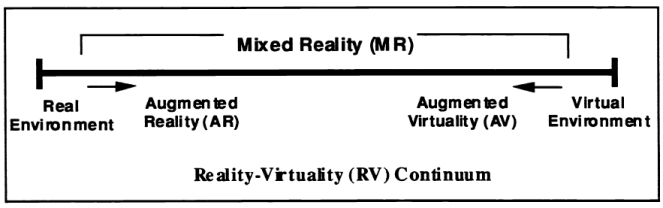

Figure 1. (Milgram et al., 1995)

Lack of Consistency in AR

AR is a new and emerging technology, and as such, its design is inconsistent. User interface (UI) elements and interactions that contribute to the user experience have not been standardized and differ across applications and hardware. Some applications are cluttered with text, whereas others are focused on the look of the holograms and provide little context. Some applications separate menu controls from the AR holograms to the point that it is difficult to understand what is being controlled, whereas others integrate them together with novel gestures that can be difficult to learn. Other applications that are used across platforms, such as a mobile phone and head-mounted display (HMD) like the Microsoft HoloLens, have entirely different controls, experiences, and gameplay elements. This inconsistency can confuse users, make it difficult to learn how to use AR technology, weaken the usability, and lessen the usage of the application.

What is Usability?

Usability has been defined as, "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction" (ISO 9241-11, 1998). The usability of a product references the user's experience with it – how well it preforms its use (effectiveness), how quickly or easily it can be used (efficiency), and how the user feels about using the product (satisfaction).

Usability is often assessed using questionnaires, usability testing, or usability heuristics. Subjective questionnaires such as the System Usability Scale (SUS) (Brooke, 1996) or the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008) can be used to gather users' perceptions about a product. Validated questionnaires are useful because they are quick to administer and can be used to compare multiple products or versions of products against each other. However, questionnaires give practitioners a very limited amount of information. These state how users rated a product, very well or very poorly, but questionnaires do not always give insights to why users gave the product their ratings.

Usability Testing

Usability testing requires users to complete tasks with a product while a researcher observes the user's behavior and notes any comments that users voice. Often, metrics such as time spent on task, task completion, number and types of errors, difficulty, confidence ratings, and first actions (or clicks) are gathered. These metrics and user comments give practitioners more feedback about why a product was given its rating, as it can be narrowed down by task, error types, and what users had to say about the product.

Usability Heuristics

Usability heuristics are lists of design guidelines that practitioners can use to evaluate products either during or after the development process. The main benefits of heuristic evaluations are that they are quicker and more cost effective than other methods to assess usability, such as usability testing (Nielsen, 1993). It is difficult to identify all of the usability problems of a product by just using heuristic evaluations, however it is useful at catching many problems that could impede users' interactions and affect their thoughts and experiences with the product. The most commonly used heuristics are Nielsen's 10 Usability Heuristics for User Interface Design (Nielsen & Molich, 1990) and Schneiderman's Eight Golden Rules of Interface Design (Schneiderman & Plaisant, 2004).

PRACTICE INNOVATION

This paper will describe both usability testing and heuristic evaluation as methods of evaluating the usability of AR applications. We will describe the challenges of these methods as they apply to AR usability research, how to overcome such challenges, and suggest future research that is needed. These challenges and recommendations have been selected based on a review of literature, as well as the authors' experience completing research projects related to AR usability testing and AR heuristic evaluations. Table 1 summarizes these key challenges and possible solutions.

PRACTICE APPLICATION

AR Usability Testing

Observing a user's screen. Some of the most common metrics gathered during usability testing are time-on-task, task success, type of errors, and number of errors (Tullis & Albert, 2013). These can be collected manually by observing the participant and noting down their behaviors or by using an automatic data collection tool. Observing a participant's screen when they are using an AR application can be tricky. Users are likely to walk around the environment instead of staying in one location, requiring an observer to move with them. The AR application may also be used with an HMD, making it impossible to see the user's screen without mirroring or sharing the screen to a separate device.

To mirror a user's screen to a TV or laptop, practitioners may need to download specialized screen casting programs, such as a Microsoft HoloLens companion tool, or buy hardware such as Miracast or Chromecast. Practitioners also can share the user's screen using Zoom or Microsoft Teams for AR applications that are being used on mobile devices. However, this will cause the device to use more processing power and can result in performance issues in the AR application. This can have an effect on how the user feels about the application, as slow response times have been shown to frustrate users (Nielsen, 1993). If screen sharing with the AR applications that are being evaluated causes a delay in response time or other glitches, it may be best to record the user's screen and review the video at a later time to collect this observational data. It is best to test the application and screen sharing method with all AR devices that are being tested before collecting any data.

Novelty. AR is a new and emerging technology that many people do not yet interact with daily. Unless the practitioner controls for past AR usage and only selects participants, who have used AR before, it is likely that many of the participants have rarely or never used AR or the specific AR device before the study. It's important to provide participants with a tutorial or training exercise so they can learn the controls of the device itself before evaluating the application.

Novelty also plays a role in how satisfied or engaged users feel about a product. If a product, like AR, is new to participants, they may simply rate it more attractive, engaging, and satisfying despite their performance. As they use the product more, the novelty wears off and could cause a decrease in user perception ratings. Practitioners can work this into their study plans by including time to practice using the AR device in-between training and test sessions.

Consistency. Consistency is crucial when conducting usability tests with AR. Lighting differences can affect the legibility of on-screen text. The amount of space in the room can affect how and where users place and interact with 3D holograms. Users may feel more comfortable to walk around 3D holograms and view them from all angles if they are in a spacious room, but not in a smaller area.

However, practitioners may want to control from where a user is viewing a hologram. For example, we conducted a text entry study with the Microsoft HoloLens Generation 1 and wanted to keep users' distance away from the text entry screen consistent as that could have an effect on their experience typing (Derby et al., 2019). As a result, we had users sit while using the application to avoid having them walk closer to or further away from the virtual screen. Practitioners should consider how consistent the

environment must be based on their research test questions and how best to balance that with mimicking natural behavior.

AR Usability Heuristic Evaluations

Using General Usability Heuristics for Evaluating AR. A practitioner may decide to conduct a usability heuristic evaluation with an AR application instead of a usability test. As mentioned, the most popular heuristics are Nielsen's 10 Usability Heuristics for User Interface Design (Nielsen & Molich, 1990) and Schneiderman's Eight Golden Rules of Interface Design (Schneiderman & Plaisant, 2004). However, these heuristics do not address specific aspects of the AR experience that could lessen the usability of the device or application. These include aspects such as comfort, spatial tracking, learning novel interaction methods, and privacy (Kourouthanassis et al., 2013; Dünser et al., 2007).

For example, the Insight Heart app developed by ANIMA RES is an AR medical education app designed for students, physicians, and patients (ANIMA RES, 2017). This app uses 3D models, animations, and user data to visualize the human heart as well as teach about different structures within the heart and four heart conditions. This application is available on mobile devices (iOS and Android), and head-mounted displays (HMDs) such as the HoloLens and Magic Leap. The mobile interface is shown in Figure 2. Some aspects of this app can be evaluated using Nielsen's 10 Usability Heuristics or Schneiderman's 8 Golden Rules. The app offers auditory and visual feedback about how the user should scan their environment to ensure that the 3D models will be placed properly (Visibility of System Status, and Error Prevention), includes UI menu items that are easily recognizable and provide affordance (Recognition Rather than Recall, and Reduce Short-Term Memory load), but fails to support undo and redo options (User Control and Freedom).

However, other aspects about this app are not addressed by these heuristics. The mobile version of Insight Heart asks the user if they would like to communicate data about their own heart rate to the app by connecting to a smartwatch. It is not always clear how the data will be stored and who will have access to it, resulting in users feeling uncomfortable using this aspect of the app. Another aspect that is not addressed is comfort. Interacting with this app by either holding up a mobile device or wearing a heavy HMD for a lengthy amount of time could cause the user to feel fatigued or strained, resulting in a negative experience or deciding to stop using the app. Jitter and lag are also concepts that may frustrate users, but are not addressed in the conventional heuristics. AR-specific usability heuristics must be used so that practitioners can create effective, efficient, and enjoyable AR experiences in a way that is quick and cost effective.

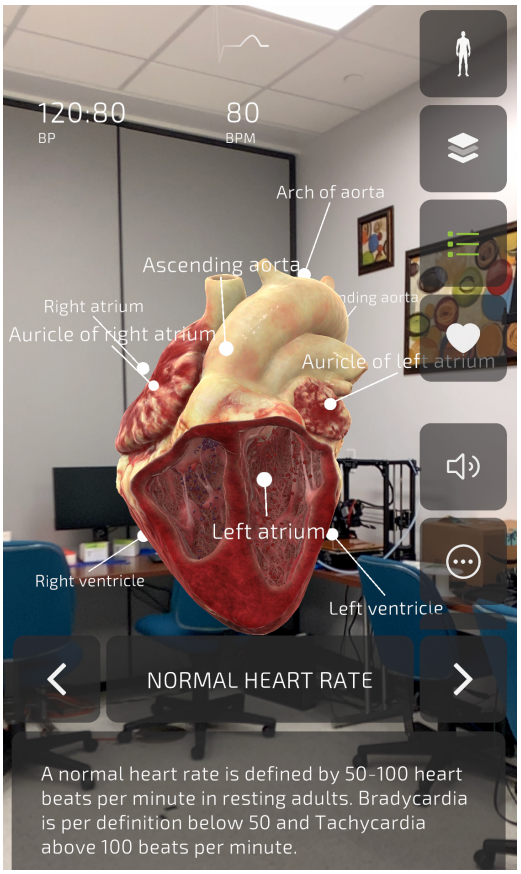

Figure 2. Insight Heart Mobile Interface.

Using AR Specific Usability Heuristics and Guidelines. In a review of the literature, we have found nine heuristic lists, usability principles, and guidelines that have been developed specifically for AR applications and devices (Ko et al., 2013; Franklin et al., 2014; de Paiva Guimarães & Martins, 2014; Gale et al., 2015; Kalalahti, 2015; Santos et al., 2016; Endsley et al., 2017; Aultman et al., 2018; Liang, 2018). Only three of which were validated (Ko et al., 2013; de Paiva Guimarães & Martins, 2014; Liang, 2018). These heuristic lists have added aspects such as user safety and comfort (Ko et al., 2013; Franklin et al., 2014; Gale et al., 2015; Kalalahti, 2015; Endsley et al., 2017), hardware capabilities (Ko et al., 2013; de Paiva Guimarães & Martins, 2014; Endsley et al., 2017), collaboration (Franklin et al., 2014), integrating virtual elements onto the physical environment (Ko et al., 2013; Franklin et al., 2014; Gale et al., 2015; Kalalahti, 2015; Endsley et al., 2017), dealing with interruptions from the physical environment (Ko et al., 2013, Gale et al., 2015), privacy (Franklin et al., 2014), and learnability (Ko et al., 2013; de Paiva Guimarães & Martins, 2014).

These heuristic sets allow practitioners to examine AR applications and devices more closely than they would with Nielsen's 10 or Schneiderman's Eight heuristic sets. However, these heuristics can be too specific for general use. For example, some of these are designed to evaluate AR smartphone apps (Ko et al., 2013; Santos et al., 2016;

Aultman et al., 2018), wearable AR (Gale et al., 2015), marker-based AR (Guimães & Martins, 2014), collaborative systems (Franklin et al., 2014), AR games (Aultman et al., 2018), or AR specifically targeted towards older adult users (Liang, 2018). This can make it difficult to compare results between devices or applications, as the heuristic set may work well for a mobile device, but not a wearable device like an HMD. AR specific heuristics that can be generalized across different devices and applications still need to be established and validated.

DISCUSSION

Augmented reality (AR) is a new and emerging technology that could benefit from practitioners' analysis of the usability of its applications. Two methods through which this could be done are usability testing and heuristic evaluations. However, because this technology blends the virtual and physical worlds together, practitioners need to keep in mind certain challenges that they may face when evaluating this technology and how to overcome those challenges. Important aspects to note are both the physical and virtual environment set-up, the capabilities of the current version of hardware that supports AR applications, user comfort and safety, and how the novelty of such a technology could affect subjective measures. By keeping all of this in mind, results on the usability of such applications and devices could give insights to how they can be designed to be more efficient, effective, and enjoyable to use.

PRACTITIONER TAKE-AWAYS

- Usability testing and usability heuristics are useful methods used to assess the effectiveness, efficiency, and satisfaction of an AR application or product.
- When completing usability tests with AR devices or applications, practitioners need to consider how note takers will observe the users screen to collect metrics on performance, the environment that the user is in, and how novelty plays a role in subjective evaluations.
- General usability heuristics such as Nielsen's 10 Usability Heuristics for User Interface Design and Schneiderman's Eight Golden Rules of Interface Design do not assess many of the aspects that affect the usability of Augmented Reality (AR) devices and applications. These include concepts such as comfort, safety, accounting for hardware capabilities, and privacy. Practitioners need to keep these aspects of AR in mind when choosing a usability heuristic set to use.

Method	Challenge	Solution(s)
Usability Testing	The act of observing a user's screen to collect data may cause the AR application to be less responsive or crash.	• Screen share using native software (ChromeCast, MiraCast, etc.). • Screen share using programs such as Zoom or Microsoft Teams. • Test screen sharing ability before testing users. • Record the user's screen instead of streaming it live.
Usability Testing	AR may be a new experience for users.	• Ensure proper tutorials and practice is given so users can get used to interacting with AR. • Keep in mind that a novelty effect may affect subjective user perception ratings such as attractiveness, engagement, and satisfaction when choosing metrics and reporting results. • Include practice time between training and test sessions to allow novelty effects to decay.
Usability Testing	The consistency of the testing environment, both virtual and physical, plays a large role.	• Ensure that the physical environment is kept consistent (e.g., adequate lighting, spacious room, decide if you want the user to sit or stand when interacting with the environment).
Usability Heuristics	Traditional heuristic sets such as Nielsen's 10 or Schneiderman's Eight do not address specific aspects of the AR experience.	• Choose heuristic sets that address aspects such as comfort, user safety, hardware capabilities, collaboration, privacy, integrating virtual and physical objects, dealing with interruptions, etc.
Usability Heuristics	Many AR-specific heuristic sets are too specific (e.g., are meant for only mobile devices, games, or specialized populations).	• Validated and widely accepted AR heuristics that can be generalized across different devices, applications, and populations should be established in future research.

Table 1. AR Usability Testing and Heuristics Challenges and Solutions.

REFERENCES

ANIMA RES. (2017). INSIGHT HEART – The human heart expedition. Retrieved from https://animares.com/portfolio/insight-heart.

Aultman, A., Dowie, S., & Hamid, N. A. (2018, April). Design Heuristics for Mobile Augmented Reality Game User Interfaces. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-5).

Barsom, E. Z., Graafland, M., & Schijven, M. P. (2016). Systematic review on the effectiveness of augmented reality applications in medical training. *Surgical endoscopy*, *30*(10), 4174-4183.

Brooke, J. (1996). Sus: a "quick and dirty" usability scale. *Usability evaluation in industry*, pp. 189-193.

Derby, J. L., Rarick, C. T., & Chaparro, B. S. (2019, November). Text Input Performance with a Mixed Reality Head-Mounted Display (HMD). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 1476-1480). Sage CA: Los Angeles, CA: SAGE Publications.

Dünser, A., Grasset, R., Seichter, H., & Billinghurst, M. (2007). Applying HCI principles to AR systems design.

Endsley, T. C., Sprehn, K. A., Brill, R. M., Ryan, K. J., Vincent, E. C., & Martin, J. M. (2017, September). Augmented Reality design heuristics: Designing for dynamic interactions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 2100-2104). Sage CA: Los Angeles, CA: SAGE Publications.

Franklin, F., Breyer, F., & Kelner, J. (2014, May). Usability heuristics for collaborative augmented reality remote systems. In *2014 XVI Symposium on Virtual and Augmented Reality* (pp. 53-62). IEEE.

Gale, N., Mirza-Babaei, P., & Pedersen, I. (2015, October). Heuristic guidelines for playful wearable augmented reality applications. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (pp. 529-534).

Google LLC. (2020). Google AR & VR. Retrieved from arvr.google.com/ar/.

ISO 9241-11. (1998). 1998 - Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. Geneva, Switzerland: ISO.

Kalalahti, J. (2015). Developing usability evaluation heuristics for augmented reality applications.

Ko, S. M., Chang, W. S., & Ji, Y. G. (2013). Usability principles for augmented reality applications in a smartphone environment. *International Journal of Human-Computer Interaction*, *29*(8), pp. 501-515.

Kourouthanassis, P. E., Boletsis, C., & Lekakos, G. (2015). Demystifying the design of mobile augmented reality applications. *Multimedia Tools and Applications*, *74*(3), 1045-1066.

Liang, S. (2018). *Establishing Design Principles for Augmented Reality for Older Adults* (Doctoral dissertation, Sheffield Hallam University).

Laugwitz, B., Schrepp, M. & Held, T. (2008). Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.

Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995, December). Augmented reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and telepresence technologies* (Vol. 2351, pp. 282-292). International Society for Optics and Photonics.

Niantic. (2020). Pokémon GO. Retrieved from https://pokemongolive.com/en/.

Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256).

Nielsen, J. (1993). Chapter 5: Usability Heuristics. In, *Usability Engineering* (pp. 115-164). Academic Press, Inc.

de Paiva Guimarães, M., & Martins, V. F. (2014, May). A checklist to evaluate augmented reality applications. In *2014 XVI Symposium on Virtual and Augmented Reality* (pp. 45-52). IEEE.

Santos, C., Miranda, B., Araujo, T., Carneiro, N., Marques, A., Mota, M., ... & Meiguins, B. (2016, July). Guidelines for graphical user interface design in mobile augmented reality applications. In *International Conference on Virtual, Augmented and Mixed Reality* (pp. 71-80). Springer, Cham.

Shneiderman, B. & Plaisant, C., (2004). *Designing the user interface: Strategies for effective Human-Computer Interaction* (4th ed.). Boston, MA: Addison Wesley.

Snap, Inc. (2019). Lens Studio. Retrieved from lensstudio.snapchat.com.

Sutherland, I. E. (1968, December). A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I* (pp. 757-764).

Tullis, T., & Albert, B. (2013). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics* (2nd ed.). Elsevier Inc.

Vilkina, M. V., & Klimovets, O. V. (2019, May). Augmented reality as marketing strategy in the global competition. In *Institute of Scientific Communications Conference* (pp. 54-60). Springer, Cham.