Evaluating Augmented Reality Systems

Chapter · July 2011

DOI: 10.1007/978-1-4614-0064-6_13

CITATIONS
36

READS
4,257

2 authors:

Andreas Duenser
The Commonwealth Scientific and Industrial Research Organisation
86 PUBLICATIONS **1,987** CITATIONS
SEE PROFILE

Mark Billinghurst
University of South Australia
548 PUBLICATIONS **14,963** CITATIONS
SEE PROFILE

Chapter 1

EVALUATING AUGMENTED REALITY SYSTEMS

Andreas Dünser and Mark Billinghurst
The Human Interface Technology Laboratory
New Zealand (HIT Lab NZ)
The University of Canterbury, Christchurch, New Zealand

1. Introduction

Augmented Reality (AR) research has a long history. Starting from the first early prototypes over forty years ago [1], AR developed into an established field of study in the 1990s and attracted interest from the larger academic community. Most recently, AR technologies also have found their way into a wide range of industrial and commercial applications.

As in most emerging technological fields, AR researchers and developers have had to solve many technological issues to create usable AR applications, such as developing tracking and display systems, authoring tools, and input devices. However, as the field matures and more applications are developed, evaluating these systems with end users will become more important. So far the amount of AR systems formally evaluated is rather small [2]. For example, literature surveys of user evaluation in AR have found that only around 8% of published AR research papers include formal evaluations [3-4]. One reason for this may be the lack of suitable methods for evaluating AR interfaces. In this chapter we review previous AR studies to give an overview of methods that have successfully been used to evaluate AR systems.

Focusing on user-driven design is an increasingly important aspect of AR system development. As more real world applications are created developers should adapt a strong user-centered design approach and evaluate the systems with actual users. This is an important step in order to bring the technology out of the research labs and into people's everyday lives. It not only helps provide more insight into the actual usage of new AR systems but also to refine their designs.

Different stages of system development call for a variety of evaluation strategies. However, not all types of evaluations test the systems' usability. Depending on the actual research goal, a user test certainly can help to shed more light on the system's usability, but researchers often have other research questions. In this chapter the term 'user evaluation' is used in a broader sense to reflect user-based studies that have a variety of different goals.

Our aim is to present several issues with AR system evaluation, and so provide researchers with a resource to help them plan and design their evaluations. In this chapter, first the role of evaluation will be described, including a review of work on user and non-user based evaluation in HCI research, and in particular for AR.

Then common evaluation techniques that have been used in AR research are examined. We close with a discussion of the issues raised in this chapter.

2. The Role of Usability Testing and User Evaluations

User evaluations are commonly called 'Usability tests'. However, this term can create some confusion. According to Nielsen usability is associated with five usability attributes: *Learnability*, *Efficiency*, *Memorability*, *Errors*, and *Satisfaction*, and usability testing involves defining a representative set of tasks against which these attributes can be measured [5]. "Usability testing is an approach that emphasizes the property of being usable […]"[6], therefore not every evaluation of an AR application is necessarily a usability test.

For example, user evaluations can be used to compare a new interaction technique with regards to user efficiency or accuracy, study user behavior and interaction with a new prototype, or study how a new system supports collaboration, etc. Thus there are many different aims that may be pursued: Is the goal to test the 'usability' of a prototype, to compare how the new technique performs compared to a benchmark system, or to check whether performance with a new interaction technique can be predicted or explained by certain concepts (e.g. Fitt's Law [7])? Usability might not always be the main reason for conducting a user evaluation.

Different evaluation techniques are used at various points in the development cycle. Techniques that are mostly employed in the early stages, such as the think aloud method or heuristic evaluation, are generally concerned with the systems' usability. However, a test conducted after the prototype has been finished is not necessarily a usability test, and may be a verification test instead [8]. Using methods developed for usability testing does not mean that one is performing a usability test. Therefore it is essential to clarify the purpose of an evaluation study.

Small informal tests can be a very valuable way to quickly uncover usability and design problems [9]. They allow for rapid iterative design input and help to ensure that the development is on track. However, they do not provide reliable results that can be generalized. We are mainly interested in finding generalizable results that can answer research questions. Uncovering usability problems of an AR prototype can be one part of a research project but generally is not the final research goal.

Greenberg and Buxton argue that evaluation can be harmful or ineffective if done naively by rule rather than by thought [10]. Academia and end-user testing often have different goals and hence validating a scientific prototype does not necessarily tell us how it is going to be used in everyday life. They also stress that the choice of evaluation method must arise from an actual problem or a research question. The inappropriate application of user evaluation can give meaningless or trivial results, destroy valuable ideas early in the design process, promote poor ideas, or give wrong ideas about how designs will be adapted by end-users.

Thus evaluating new designs with users can be a very powerful tool in the research process but it should only be done if it employed correctly. User studies can only add value if they are rigorously designed, carried out, and analyzed, and if they help to answer a meaningful research question.

3. AR System Evaluation

There are several challenges for researchers trying to evaluate AR user interfaces. Research on AR systems is still relatively young so there is limited experience with a variety of evaluation design factors. We will discuss some of these challenges including difficulties in applying existing evaluation guidelines, a lack of knowledge about end-users, and the huge variety of software and hardware implementations.

3.1 Applying traditional evaluation approaches

Many authors agree that researchers in emerging interface fields such as Virtual Reality (VR) or AR cannot rely solely on design guidelines for traditional user interfaces [11] [12] [2] [13] [14]. New interfaces afford interaction techniques that are often very different from standard WIMP (window, icon, menu, pointing device) based interfaces, but most available guidelines were specifically developed for WIMP interfaces. Sutcliffe and Kaur [15] argue that although traditional usability evaluation methods may be able to uncover some problems with novel interfaces none of the current methods really fit the specific needs of such interfaces.
Stanney and colleagues [14] list limitations of traditional usability methods for assessing virtual environments:

- Traditional point and click interactions vs. multidimensional object selection and manipulation in 3D space
- Multimodal system output (visual, auditory, haptic) is not comprehensively addressed by traditional methods
- Assessing presence and after-effects not covered by traditional methods
- Traditional performance measures (time, accuracy) do not always comprehensively characterize VE system interaction
- Lack of methods to assess collaboration in the same environment

Such limitations apply to AR interfaces as well because they share many characteristics with virtual environments. WIMP inspired heuristics do not always fit the design needs of AR systems. For example Nielsen's well known usability heuristics [16] do not cover issues relating to locating, selecting and manipulating objects in 3D space. Similarly, input and output modalities can be radically different for AR interfaces, requiring different evaluation approaches.

3.2 One off prototypes

Developing a general evaluation framework for AR systems can be rather challenging. Although we have made great progress in AR research many of the developed systems are one off prototypes. Researchers often design and create a prototype as a proof of concept, test it, and then move on to the next problem. This leaves us with a collection of interesting ideas and artifacts that generally do not share many common aspects. This complicates the development of common design and evaluation guidelines.

In contrast, WIMP interfaces share many characteristics and therefore have design features in common. These interfaces are more established and new developments are more likely feature incremental changes rather than radical ones [17] which is often the case with new AR developments.

3.3 Who are the users?

With many technological developments the key question is who are the end users and what are their needs [18]. This is especially true when products are developed for the consumer market. Until recently there have been very few AR-based consumer products. This changed when different mobile and desktop AR applications were brought to market, but few of these have been formally evaluated with users.
If developments are technology driven rather than user-need driven it is not clear who will actually use the AR application. We are therefore often presented with a solution which solves no real problem. If we do not have a clear idea about whom the end users will be we do not know with whom we should evaluate the system in order to get representative results. This raises the question about whether a system evaluation would be meaningful in the first place. If it is not obvious how and by whom the system will be used, it is rather difficult to design a meaningful system and evaluation plan. The users may be novices, experts, casual users, frequent users, children, adults, elderly, and so on. All these characteristics affect the ways in which interaction is designed [6].

3.4 Huge variety of implementations

Compared to WIMP interfaces that share basic common properties, AR interfaces can be much more diverse. Although we can follow certain definitions, such as Azuma's [19], to determine what constitutes an AR system, this still leaves us with many different types of systems. The digital augmentation can include different senses (sight, hearing, touch, etc.), be realized with an array of different types of input and output hardware (mobile, Head Mounted Display (HMD), desktop, etc.), and allow for a variety of interaction techniques. This makes it challenging to define a common and comprehensive set of evaluation techniques.
One solution could be to create a framework for AR systems and/or interaction as a basis for the development of design and evaluation guidelines. The challenge will be to find an acceptable abstraction level that is necessary to create such a comprehensive framework that would still allow creating practical guidelines.
Another solution would be to narrow down the field of interest to an extent that it permits the definition of common guidelines. For example, creating guidelines for mobile phone AR systems. This would lead to, different sets of guidelines for different kinds of AR systems that will share certain characteristics.

4. Expert and Guideline Based Evaluations

In the research literature we can find various classifications of usability evaluation techniques e.g. by Bowman, Gabbard and Hix [11]. For simplicity in this chapter

we separate evaluation approaches in two rather broad categories; evaluation techniques that do not involve users, and those that do involve users. In this section methods are discussed that do not require users, and instead use experts.

Heuristic evaluation is a usability testing method in which experts are asked to comment on and evaluate an interface design using a set of design guidelines and principles [16]. After familiarizing themselves with the system the evaluators carry out certain tasks and note the interface problems they found. A limited set of heuristics is then used to interpret and classify these problems [13].

Some attempts have been made to develop guidelines and heuristics for the evaluation of VE [13, 14, 20]. These research efforts generally produce very long lists of heuristics that have been extracted from research literature. This reflects some of the problems discussed above. The huge variety of implementation possibilities (hard-/software, interaction devices, etc.) leads to a huge number of guidelines.

Sutcliffe and Gault [13] give a very concise list of 12 heuristics. They include a technology audit stage before the evaluation in order to filter the heuristics and make sure that only relevant ones are applied. They found that some evaluators had difficulty in interpreting some heuristics. To overcome this, the authors suggest giving good and bad examples for each item to better illustrate the heuristics. Some of these heuristics can be applied to AR systems with slight adaptations. 'Natural engagement', 'Compatibility with the user's task and domain', 'Realistic feedback', 'Support for learning', 'Clear turn-taking', and 'Consistent departures' are mostly applicable to AR environments as well. Others heuristics such as 'Natural expression of action', 'Close coordination of action and representation', 'Faithful viewpoints', 'Navigation and orientation support', 'Clear entry and exit points', and 'Sense of presence' might be less applicable for AR because they are mainly concerned with the users' representation or navigation in the VE.

Stanney et al. [14] have created a computerized system to assist in the evaluation of virtual environments. MAUVE (Multi-criteria Assessment of Usability for Virtual Environments) aims at structuring the evaluation process and to help managing the huge amount of guidelines that aid in the virtual environment (VE) assessment (the author's list four pages of design considerations/guidelines). The system is based on a hierarchical framework illustrated in Figure 1.

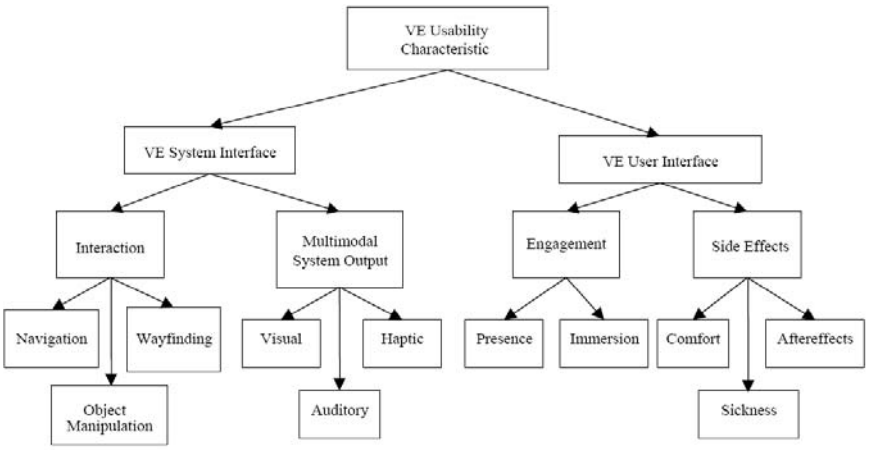

Figure 1. Usability criteria assessed by MAUVE [14]

Sutcliffe and Kaur [20] present a walkthrough method for VE user interfaces designed desktop VR systems. This is sensible because it limits the implementation options and therefore the amount of potential guidelines. It also makes the task of defining an interaction framework and creating a concise list of appropriate guidelines more manageable. They specify models of interaction as a basis for the guidelines based on goal-oriented task action, exploration and navigation in VE, and interaction in response to system initiatives. The authors acknowledge that walkthroughs may be more efficient in finding usability problems rather than in helping finding design solutions. They also point out the problem of context and that some general design principles may be inapplicable in different contexts.

Some of these guidelines might be applicable to AR environments as well as guidelines on object selection and manipulation (e.g. Input devices should be easy to use; Object selection points should be obvious, and it should be easy to select multiple objects), multimodal output (e.g. Visual, auditory, and/or haptic should have high frame rate and low latency, and be seamlessly integrated into user activity), and side-effects (e.g. system should be comfortable for long-term use).

One of the few efforts to define design and evaluation guidelines for AR systems was presented by Gabbard (2001). His approach was to collect and synthesize information from many different sources including AR specific research and more general VE based research. The result of this research is an extensive list of 69 statements or guidelines, in several categories, including:

- VE Users and User tasks
  - VE Users (e.g. “Support users with varying degrees of domain knowledge.”)
  - VE user tasks (e.g. “Design interaction mechanisms and methods to support user performance of serial tasks and tasks sequences.”)
  - Navigation and Locomotion (e.g. “Support appropriate types of user navigation, facilitate user acquisition of survey knowledge.”)

  - Object selection (e.g. “Strive for body-centered interaction. Support multimodal interaction.”)
  - Object manipulation (e.g. “Support two-handed interaction (especially for manipulation-based tasks).”)
- The Virtual Model
  - User Representation and Presentation (e.g. “For AR-based social environments (e.g., games), allow users to create, present, and customize private and group-wide information.”)
  - VE Agent Representation and Presentation (e.g. “Include agents that are relevant to user tasks and goals.”)
  - Virtual Surrounding and Setting (e.g. “When possible, determine occlusion, dynamically, in real-time (i.e., at every graphics frame).”)
  - VE System and Application Information (e.g. “Strive to maintain interface consistency across applications”.)
- VE User Interface Input Mechanisms
  - Tracking User Location and Orientation (e.g. “Consider using a Kalman Filter in head tracking data to smooth the motion and decrease lag.”)
  - Data Gloves and Gesture Recognition (e.g. “Avoid gesture in abstract 3D spaces; instead use relative gesturing.”)
  - Speech Recognition and Natural Language (e.g. “Allow users to edit, remove, and extract or save annotations.”)
- VE User Interface Presentation Components
  - Visual Feedback -- Graphical Presentation (e.g. “Timing and responsiveness of an AR system are crucial elements (e.g., effect user performance).”)

Most of the guidelines listed in Gabbard’s work were taken from papers on VE systems and are not necessarily specific to AR systems (such as guidelines on navigation). This again illustrates the challenge of creating a succinct list of AR design guidelines that is easy to apply by practitioners and researchers.

Generally heuristics do not seem to be used very often to evaluate AR systems. Apart from the limited practicality of using extensive lists of heuristics and the limited numbers of experts to do such evaluations, guideline based evaluation is more of a usability evaluation tool rather than a research tool. By using these techniques experts might uncover usability problems with a system, but this hardly permits researchers to answer major research questions. From a research perspective these heuristics might be useful in informing prototype design. Following design guidelines can help reduce time for prototype development and develop more stable systems for user testing. On the other hand, as a testing tool they are more important for industry rather than for researchers. If the goal is to develop an easy to use system for end-users, expert and guideline-based evaluation can be very valuable tools. One of the few AR studies that employed expert based heuristic evaluation was reported by Hix et al. [21] and is discussed in section 5.2.

5. User based evaluation

Fjeld [18] argues that researchers still have to find appropriate ways to measure effectiveness and efficiency in AR applications and to define what kind of tasks and with what kind of tools the usability of AR systems can be tested.
Evaluation approaches used in traditional HCI are often applied to AR research, but evaluating AR systems with users sometimes requires slightly different approaches than evaluating traditional GUI based systems. Some issues of evaluating novel user interfaces are discussed by Bowman et al. [11], such as the physical environment in which the interface is used (including interaction issues), evaluator issues (and how they deal with the complexities of VEs), and user issues. For example most users of new systems will be novices because only a few people (e.g. the developers) can be considered as experts. This may produce high variability in the obtained results and therefore require a high number of study participants.
The design space for novel interfaces is generally very large and often more than two competing solutions have to be compared. This requires more complicated study and analysis designs that again call for higher participant numbers. Furthermore, many effects related to simulated environments such as presence or simulator sickness require new methods not considered by GUI evaluation techniques.
Evaluating AR interfaces frequently focuses on somewhat different issues than traditional GUI / WIMP evaluation. With AR systems increasing the user's effectiveness and efficiency are not always the primary goals. Many WIMP based systems are designed to support users in accomplishing specific tasks effectively (e.g. office work, word processing etc.). While some AR systems pursue similar goals (e.g. AR systems for engineering), most seem to focus more on providing a novel user experience that require different evaluation techniques [22] [23] [24].

5.1 AR Evaluation types and methods

In previous work [4] we summarized the different evaluation techniques that have been applied in AR user studies. We reviewed evaluation types based on typical tasks that are performed in AR applications and analyzed the evaluation techniques that have been used. In the rest of this section we present these categories and discuss them with examples taken from scientific literature.

5.1.1 Evaluation types typically used in AR user evaluations:

Based on work conducted by Swan and Gabbard [3] and from our own survey [4] most AR user evaluations fit into one of four categories:

(1) Experiments that study human perception and cognition with low-level tasks.

(2) Experiments that examine user task performance.

(3) Experiments that examine collaboration between users.

(4) System usability, system design evaluation.

5.1.2 Evaluation methods typically used in AR user evaluations:
A rough distinction can be made between quantitative methods, qualitative methods, non-user based usability evaluation methods, and informal methods.

(1) Objective measurements
These should produce a reliable and repeatable assignment of numbers to quantitative observations. They can be taken automatically or by an experimenter. Typical measures include times (e.g. task completion times), accuracy (e.g. error rates), user or object position, or test scores, etc.

(2) Subjective measurements
These rely on the subjective judgment of people and include questionnaires, ratings, rankings, or judgments (e.g. depth judgment).

(3) Qualitative analysis
Qualitative analysis is not concerned with putting results in numbers. Data is gathered through structured observations (direct observation, video analysis) or interviews (structured, unstructured).

(4) Non User-Based Usability evaluation techniques
This includes non user-based evaluations techniques such as cognitive walkthroughs or heuristic evaluations as well as techniques that involve people who are not end-users (e.g. expert-based usability evaluations).

(5) Informal testing
Many published AR papers only report on informal user observations or feedback (e.g. gathered during demonstrations). It is surprising that reporting such limited findings still seems to be very common and accepted in AR contexts. By contrast, in CHI publications informal evaluation has almost disappeared [25].

5.2 Example AR system evaluations
The following papers provide examples of the different evaluation methods. Studies with more than one evaluation type or method are grouped according to the main focus.

Type: (1) low-level tasks – perception; method: (1) objective measurements
Gabbard and Swan [26] studied how users perceive text in outdoor AR settings. They tested several hypotheses, such as seeing text on a brick background will result in slower task performance because it is the most visually complex. They tested 24 volunteers who had to identify a pair of letters in a random arrangement (see figure 2). Independent variables included background texture, text color, drawing style, and drawing algorithm. Dependent variables were response time and error. Results showed that the participants made the least errors with a solid building background and most with a brick background. No main effects of text color were found but the participants were slower with billboard style text.

Figure 2: Experimental task used by Gabbard and Swan [26]: participants were asked to to identify the pair of identical letters in the upper block (“Vv”)

Type: (1) low-level task – perception; method: (2) subjective measurements
Knörlein et al. [27] investigated the effect of visual and haptic delays on the perception of stiffness in an AR system. Fourteen participants were asked to compress and compare two virtual springs in a visuo-haptic environment and to select the one which they perceived as stiffer.

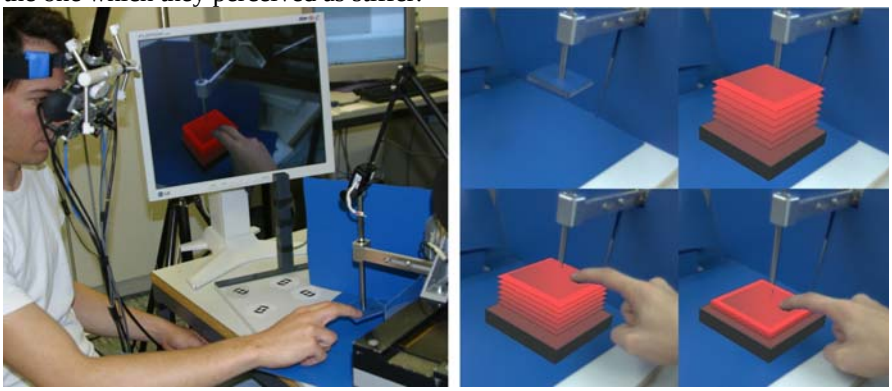

Figure 3: Study setup used by Knörlein et al. [27] to test effects of visual and haptic feedback on perception in an AR system

The metrics used in this study were ‘Point of Subjective Equality’ and ‘Just Noticeable Difference’ of the stiffness of the two springs. Results showed that delays for haptic feedback decreased perceived stiffness whereas visual delays increased

perceived stiffness. When visual and haptic delays were combined the effects leveled each other out.

Type: (2) user performance; method: (1) objective measurements
The goal of Dünser et al. [28] was to evaluate if spatial ability could be improved through training with an AR-based geometry education application. The system allows students to collaboratively construct geometry problems in 3D space. The interface consists of optical see-through HMDs and position tracked pens and tablets that allow direct manipulation of the virtual content (see figure 4).

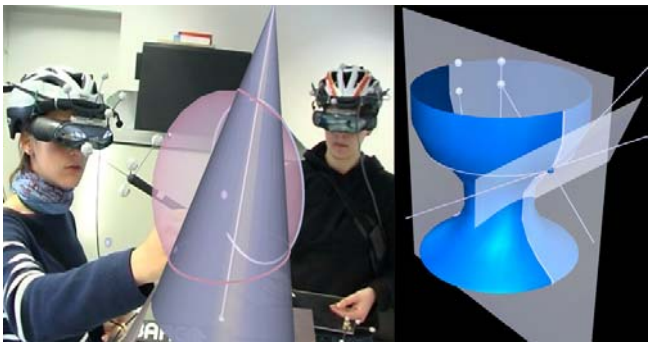

Figure 4: The Construct3D system, an AR geometry education application [28-29]

The study participants were 215 high school students, split into four groups; AR training, traditional computer-based training, control group with geometry classes and control group without geometry classes. Students in the two training groups completed six training sessions in which they worked on a set of geometry tasks. To determine training effects, four tests measuring different aspects of spatial ability were employed before and after the training sessions: Differential Aptitude Test: Space Relations (DAT:SR) [30] (paper folding task; visualization), Mental Cutting Test (MCT) [31]) (object cutting; visualization), Mental Rotation Test (MRT) [32] (speeded mental rotation), Objective Perspective Test (OPT) [33] (change of perspective; orientation). The results showed various gender specific effects but no clear evidence of the effectiveness of AR-based training.

Type: (3) user collaboration; method: (1) objective measurements (and subjective measurement)

Billinghurst et. al. [34] evaluated how different interfaces support collaboration. They compared face-to-face collaboration with AR, and projection displays. They were interested how user collaboration differs between these setups. They hypothesized that collaboration with AR technology will produce behaviors that are more like natural face-to-face collaboration than those behaviors produced by a projection based interface. Fourteen participant pairs had to solve an urban design logic puzzle and arrange nine buildings to satisfy ten rules in seven minutes.

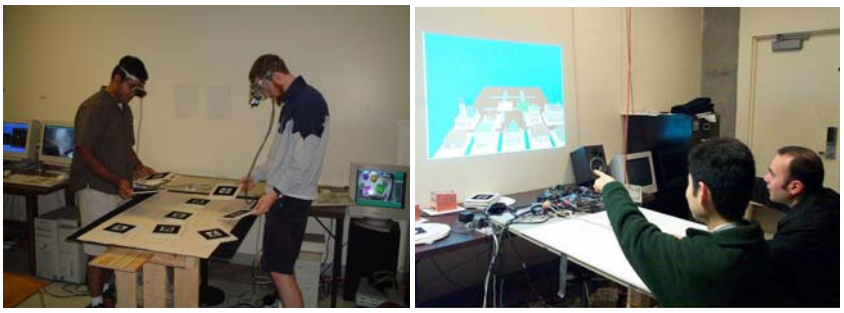

Figure 5: AR (left) and Projection (right) conditions used by Billinghurst et al. [34] to study user collaboration.

A variety of different experimental measures were used; video was captured and transcribed to analyze various communication measures, including the number of gestures, average number of words per phrase, and number turns in the conversation. Although the analysis was based on observation and video analysis, the main focus was on gathering quantities (i.e. numbers of gestures and utterances). Further measures were task completion time and questionnaires.
They found that performance with AR supported collaboration was slower than in the face-to-face and projection conditions. However, pointing and picking gesture behaviors as well as deictic speech patterns were found to be the same in AR and face-to-face and significantly different than in the projection condition. Questionnaire results indicated that users felt that it was easier to work together in the face-to-face condition. Interaction was found to be similar in AR and face-to-face cases, and much easier than in the projection condition.

Type: (2) User interaction (and collaboration); method: (3) qualitative analysis (and subjective measures)

Morrison et al. [23] conducted a qualitative analysis of user behavior and complemented this with questionnaire data. They studied user interaction and collaboration using a location-based mobile game. Twenty six participants (in pairs or teams of three) used a mobile AR interface with a paper map, while eleven participants used a traditional 2D mobile phone based map. The application was an environmental awareness game in which the participants had to solve several tasks according to some clues. Researchers followed the participants during the game and took notes, videos, and photographs which were later qualitatively analyzed. This data was complemented with data gathered through an interview as well as presence, flow, and intrinsic motivation questionnaires. The study found that the AR-based system supports collaboration, negotiating, and establishing common ground but was more cumbersome to use. With the 2D mobile phone map the participants completed the game quicker and showed less focus on the interface itself.

Figure 6: Still shots form experimenter/observer camera taken during Morrison et al.'s [23] user trials.

Type: (4) system usability, system design evaluation; method: (3) qualitative analysis

Nielsson and Johansson [35] investigated how participants experience instructions given by an AR system in a medical environment. They conducted two qualitative user studies to investigate user experience and acceptance of an instructional AR system. This was tested with eight (first study) and twelve (second study) participants respectively. The system included an HMD setup, marker tracking, and a keyboard or microphone and voice recognition for simple command input.
In the experiment the participants received AR-based instructions on how to activate and prepare medical equipment for surgery. After this they had to assemble a common medical device. The analysis was based on observations and a questionnaire. The questionnaire had an open answer format and questions on overall impression of the AR system, experienced difficulties, experienced positive aspects, what they would change in the system and whether it was possible to compare receiving AR instructions to receiving instructions from a teacher.
Results showed that the participants were able to complete tasks without assistance but they wished for more interactivity. The instructions and the AR presentation of instructions were rated positively by most users. The study also uncovered several ergonomic issues such as distraction because of marker visibility.

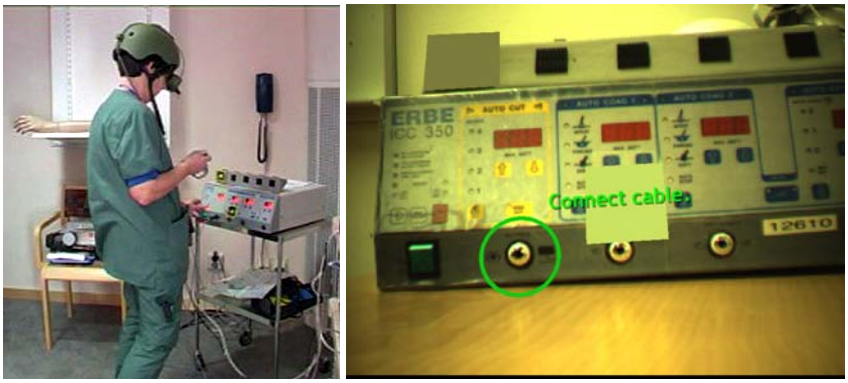

Figure 7: Study setup (left) and user view (right) of Nielsson and Johansson's instructional AR system [35]

Type: (4) system usability, system design evaluation; method: (4) usability evaluation techniques (and other methods)

Finally, Hix et al. [21] provide a nice example of how AR systems can be evaluated with a complete set of usability evaluation techniques. They discuss a model for a cost-effective usability evaluation progression, depicted in Figure 8.

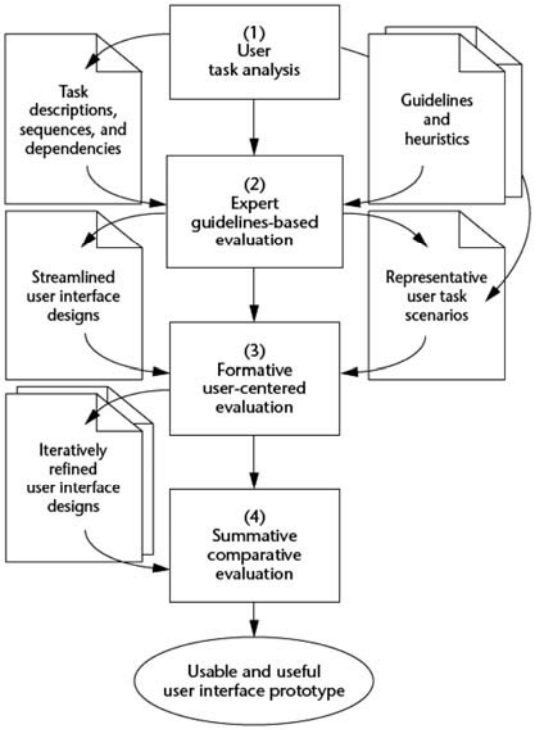

Figure 8: Evaluation methodology for VE user interaction [26]

Their paper demonstrates the practical application of this model by describing the interface design and usability evaluation of an outdoor AR system that supports information presentation and entry for situation awareness in an urban war fighting setting. They started with a domain and requirement analysis, followed by six cycles of expert evaluations in which approximately 100 mockups were designed. In a user-based evaluation perceptual issues such as line drawing style and opacity levels were studied. Eight participants had to indicate the location of a target as it moved among different buildings. Task completion time and error rates were analyzed statistically. In the following step the authors conducted what they call a formative usability evaluation. For this they created a formal set of representative tasks which five expert users had to perform. The participants had to find specific task relevant information in the outdoor environment. Analysis was qualitative

and focused on system usability issues. The authors did not conduct a summative usability evaluation so there is no available information on this last planned step.

6 Discussion and Conclusion

Many researchers agree that formal evaluation of AR interfaces is an important step in the research process. However many open questions remain. There is no agreement on which are the most appropriate evaluation methods for AR systems, and trying to find 'the best' method(s) might not be a very promising approach. The most suitable evaluation approach always depends on the questions posed.

This chapter gives an overview of the most commonly used evaluation approaches used in AR research literature. First we discussed how non user-based approaches such as heuristic evaluation have not been used very often in AR research. Several challenges complicate the process of defining a comprehensive list of heuristics and guidelines for designing and evaluating AR systems. Initial attempts have been made to derive such guidelines for VE systems. For AR systems there are even less practically available heuristic guidelines. The question is whether it is possible to develop a general set of guidelines that are applicable to a broad variety of systems. Some general heuristics can serve as a starting point, but this approach might be limiting because it will only apply to high level design issues.

Evaluations based on measuring user task performance are the most popular in AR evaluations [4]. It is not clear if this is due to a lack of more appropriate methods or if improving user task performance really is the most important research question. As argued in this paper, issues relating to user experience can be very important for many AR systems. However, there are hardly any suitable methods for measuring user experience. If user experience is of interest researchers most generally use qualitative evaluation approaches, but there are still few AR studies using these approaches. In future we should continue exploring a variety of different methods to add to the set of evaluation methods suitable for AR specific research.

Many AR environments afford various other interaction possibilities that are to a lesser extend supported by traditional computer interfaces and therefore have gained less attention in these contexts. So researchers have to search in other disciplines for appropriate methods. One example is studying interaction between users. AR interfaces offer different possibilities for collaboration but very few studies aiming at evaluating collaboration in AR environments have been published. Suitable evaluation methods for studying collaboration can be found in fields such as Computer Supported Collaborative Work.

Whether the goal of an AR system evaluation is to find usability problems or to answer a specific research question we will have to adapt current approaches to better fit the specific requirements of AR-based interfaces. In order to derive meaningful heuristics and guidelines we cannot resort to just collecting various guidelines from other scientific publications or through our own experiments. Such guidelines can only be sensibly applied if specific contexts are considered. Therefore we have to begin with setting guidelines in context through developing specific frameworks and models for AR systems and AR interaction techniques. At this stage it is not entirely obvious whether it will be possible to have a single set

of frameworks that encompass all kinds of AR systems or if there have to be separate, more context specific frameworks. These models would also help to better understand how and where to use the other evaluation methods and how we have to adapt these to be more suitable for testing AR systems.

References

1. I. Sutherland, "A Head-Mounted Three-Dimensional Display," presented at the Fall Joint Computer Conf., Am. Federation of Information Processing Soc. (AFIPS), Washington, D.C., USA, 1968.
2. A. Dünser, R. Grasset, H. Seichter, and M. Billinghurst, "Applying HCI principles to AR systems design," Charlotte, NC, USA, 2007.
3. J. E. Swan and J. L. Gabbard, "Survey of User-Based Experimentation in Augmented Reality," presented at the 1st International Conference on Virtual Reality, HCI International 2005, Las Vegas, USA, 2005.
4. A. Dünser, R. Grasset, and M. Billinghurst, "A survey of evaluation techniques used in augmented reality studies," presented at the ACM SIGGRAPH ASIA 2008 courses, Singapore, 2008.
5. J. Nielsen, Usability Engineering. San Francisco: Morgan Kaufmann, 1993.
6. H. Sharp, Y. Rogers, and J. Preece, Interaction Design: Beyond Human-Computer Interaction: Wiley, 2007.
7. P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," Journal of Experimental Psychology, vol. 47, pp. 381-391, 1954.
8. R. S. Dicks, "Mis-usability: on the uses and misuses of usability testing," presented at the Proceedings of the 20th annual international conference on Computer documentation, Toronto, Ontario, Canada, 2002.
9. J. Nielsen, Designing Web Usability. Indianapolis, IN, USA: New Rivers, 2000.
10. S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)," CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, 2008.
11. D. A. Bowman, J. L. Gabbard, and D. Hix, "A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods," Presence - Teleoperators and Virtual Environments, vol. 11, pp. 404-424, 2002.
12. C. Bach and D. L. Scapin, "Adaptation of Ergonomic Criteria to Human-Virtual Environments Interactions," presented at the INTERACT 2003.
13. A. Sutcliffe and B. Gault, "Heuristic evaluation of virtual reality applications," Interacting with Computers, vol. 16, pp. 831-849, 2004.
14. K. M. Stanney, M. Mollaghasemi, L. Reeves, R. Breaux, and D. A. Graeber, "Usability engineering of virtual environments (VEs): identifying multiple criteria that drive effective VE system design," Int. J. Hum.-Comput. Stud., vol. 58, pp. 447-481, 2003.
15. A. Sutcliffe and K. Kaur, "Evaluating the usability of virtual reality user interfaces," Behaviour and Information Technology, vol. 19, 2001.

16. J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," presented at the Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people, Seattle, Washington, United States, 1990.
17. W. Newman, "A Preliminary Analysis of the Products of HCI Research, Using Pro Forma Abstracts," presented at the CHI, Boston, MA, 1994.
18. M. Fjeld, "Introduction: Augmented Reality-Usability and Collaborative Aspects," International Journal of Human-Computer Interaction, vol. 16, p. 387 — 393, 2003.
19. R. T. Azuma, "A Survey of Augmented Reality," Presence - Teleoperators and Virtual Environments, vol. 6, pp. 355-385, 1997.
20. A. Sutcliffe and K. Kaur, "Evaluating the usability of virtual reality user interfaces," Behaviour and Information Technology, vol. 19, 2000.
21. D. Hix, J. L. Gabbard, J. E. S. II, M. A. Livingston, T. H. Höllerer, S. J. Julier, Y. Baillot, and D. Brown, "A Cost-Effective Usability Evaluation Progression for Novel Interactive Systems," presented at the Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 9 - Volume 9, 2004.
22. S. W. Gilroy, M. Cavazza, and M. Benayoun, "Using affective trajectories to describe states of flow in interactive art," presented at the Proceedings of the International Conference on Advances in Computer Enterntainment Technology, Athens, Greece, 2009.
23. A. Morrison, A. Oulasvirta, P. Peltonen, S. Lemmela, G. Jacucci, G. Reitmayr, J. Näsänen, and A. Juustila, "Like bees around the hive: a comparative study of a mobile augmented reality map," presented at the Proceedings of the 27th international conference on Human factors in computing systems, Boston, MA, USA, 2009.
24. E. Hughes, E. Smith, C. B. Stapleton, and D. E. Hughes, "Augmenting Museum Experiences with Mixed Reality," presented at the Knowledge Sharing and Collaborative Engineering, St. Thomas, US Virgin Islands, 2004.
25. L. Barkhuus and J. A. Rode, "From Mice to Men – 24 years of Evaluation in CHI," presented at the CHI, San Jose, USA, 2007.
26. J. L. Gabbard and J. E. Swan, "Usability Engineering for Augmented Reality: Employing User-Based Studies to Inform Design," IEEE Transactions on Visualization and Computer Graphics, vol. 14, pp. 513-525, 2008.
27. B. Knörlein, M. D. Luca, and M. Harders, "Influence of visual and haptic delays on stiffness perception in augmented reality," presented at the Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality, 2009.
28. A. Dünser, K. Steinbügl, H. Kaufmann, and J. Glück, "Virtual and augmented reality as spatial ability training tools," presented at the Proceedings of the 7th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI, Christchurch, New Zealand, 2006.
29. H. Kaufmann, K. Steinbügl, A. Dünser, and J. Glück, "General Training of Spatial Abilities by Geometry Education in Augmented Reality," Annual Review of CyberTherapy and Telemedicine: A Decade of VR, vol. 3, pp. 65-76, 2005.

30. CEEB College Entrance Examination Board, Special Aptitude Test in Spatial Relations MCT: CEEB, 1939.
31. G. K. Bennett, H. G. Seashore, and A. G. Wesman, Differential Aptitude Tests, Forms S and T. New York: The Psychological Corporation, 1973.
32. M. Peters, B. Laeng, K. Latham, M. Jackson, R. Zaiyouna, and C. Richardson, "A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance," Brain and Cognition, vol. 28, pp. 39-58, 1995.
33. M. Hegarty and D. Waller, "A dissociation between mental rotation and perspective-taking spatial abilities," Intelligence, vol. 32, pp. 175-191, 2004.
34. M. Billinghurst, H. Kato, K. Kiyokawa, D. Belcher, and I. Poupyrev, "Experiments with Face-To-Face Collaborative AR Interfaces," Virtual Reality, vol. 6, pp. 107-121, 2002.
35. S. Nilsson and B. Johansson, "Acceptance of augmented reality instructions in a real work setting," presented at the CHI '08 extended abstracts on Human factors in computing systems, Florence, Italy, 2008.

Index terms (alphabetically):

Qualitative analysis
Questionnaire
Subjective measurement
Usability
User
    End-user
    User performance
Virtual reality, VR
Virtual Environment, VE
Window Icon Mouse Pointing device, WIMP