Proceedings of the Hawaii International Conference on System Sciences (HICSS-37), January 2004.

A Cost-Effective Usability Evaluation Progression for Novel Interactive Systems

Deborah Hix, Ph.D.
Systems Research Center, Virginia Tech, Blacksburg, VA
hix@vt.edu

Joseph L. Gabbard, M.S.
Systems Research Center, Virginia Tech, Blacksburg, VA
jgabbard@vt.edu

J. Edward Swan II, Ph.D.
Virtual Reality Laboratory, Naval Research Laboratory, Washington, D.C.

Mark A. Livingston, Ph.D.
Virtual Reality Laboratory, Naval Research Laboratory, Washington, D.C.

Tobias H. Höllerer, Ph.D.
Department of Computer Science, University of California, Santa Barbara, CA

Simon J. Julier, Ph.D.
ITT Advanced Engineering and Sciences, Alexandria, VA

Yohan Baillot, M.S.
ITT Advanced Engineering and Sciences, Alexandria, VA

Dennis Brown, M.S.
Virtual Reality Laboratory, Naval Research Laboratory, Washington, D.C.

Abstract

This paper reports on user interface design and evaluation for a mobile, outdoor, augmented reality (AR) application. This novel system, called the Battlefield Augmented Reality System (BARS), supports information presentation and entry for situation awareness in an urban war fighting setting. To our knowledge, this is the first time extensive use of usability engineering has been systematically applied to development of a real-world AR system.

Our BARS team has applied a cost-effective progression of usability engineering activities from the very beginning of BARS development. We discuss how we first applied cycles of structured expert evaluations to BARS user interface development, employing user interface mockups representing occluded (non-visible) objects. Then we discuss how results of these evaluations informed our subsequent user-based statistical evaluations and formative evaluations, and present these evaluations and their outcomes. Finally, we discuss how and why this sequence of types of evaluation is cost-effective.

1. Introduction

For more than two decades, through our work in human-computer interaction and usability engineering, we have pursued the goals of developing, applying, and extending methods for improving the usability of interactive software applications. In particular, our work has focused on high-impact, cost-effective techniques for evaluating usability of interactive systems. By "high-impact" and "cost-effective", we mean that we have as a goal the development of methodological techniques that reduce the total life cycle cost of an interactive software application.

Usability engineering produces highly usable user interfaces that are essential to improved user experiences and productivity, as well as reduced user errors. Unfortunately, managers and developers often have the misconception that usability engineering activities *add* costs to a product's development life cycle. In fact, usability engineering can *reduce* development costs over the life of the product, by, for example, decreasing the need to add missed functionality later in the development cycle when such additions are much more expensive. The process is an integral part of interactive software development, just as are systems engineering and software engineering. Usability engineering activities can be tailored to allow individualization as needed for a specific project or product development effort.

The usability engineering process applies to any interactive system, ranging from training applications to multimedia CD-ROMs to augmented and virtual environments to simulation applications to graphical user interfaces (GUIs). The usability engineering process is flexible enough to be applied at any stage of the development life cycle, and its various activities are generalizable and adaptable across development of all interactive systems. However, just like good software engineering practices [1], early use of the process provides the best opportunity for cost savings.

In this paper, we discuss user interface design and evaluation for a mobile, outdoor, augmented reality (AR) application. This novel system, called the Battlefield Augmented Reality System (BARS), supports information presentation and entry for situation awareness when conducting urban military operations. We have systematically incorporated a cost-effective progression of usability engineering activities from the

very beginning of BARS development. Thus, this paper focuses on the specific process by which we evaluated the BARS product. Results of our usability engineering process as applied to numerous other products can be found, for example, in [4, 5, 7, 9, 10].

To our knowledge, this is the first time usability engineering has been extensively and systematically applied to the research and development process of a real-world AR system. In fact, a comprehensive literature review of 880 papers from the leading augmented reality/virtual reality conferences and publication sources showed 25 papers (less than 3%) that had any human-computer interaction discussion, and of those, only 14 (about 1.5%) reported a user-based study [15].

2. What is usability engineering?

Usability engineering is a cost-effective, user-centered process that ensures a high level of effectiveness, efficiency, and safety in complex interactive systems [6]. Figure 1 shows a simple diagram of major usability engineering activities, which include domain analysis, quantifiable user-centered requirements and metrics, conceptual and detailed user interface design, rapid prototyping, and various kinds of usability evaluations of the user interface. Usability engineering includes both design and evaluations with users; it is not typically extensive hypothesis-testing-based experimentation, but instead is structured, iterative user-centered design and evaluation applied during all phases of the interactive system development life cycle. Most extant usability engineering methods widely in use were spawned by the development of traditional desktop graphical user interfaces (GUIs).

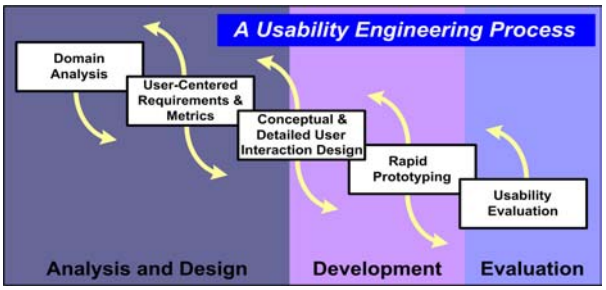

Figure 1. Typical activities performed during the usability engineering process. Although the usual flow is left-to-right from activity to activity, outward-pointing arrows indicate the substantial feedback and iterations that occurs in practice.

Since the focus of this paper is the usability engineering activity of *usability evaluation*, in the following sections we briefly explain several types of usability evaluation. These include *expert evaluation* (also sometimes called heuristic evaluation or usability inspection), *user-based statistical evaluation*, *formative evaluation*, and *summative evaluation*. These introductory and brief explanations are, of necessity, rather abstract, to introduce each type of evaluation. In Section 4, we present, very concretely, how we applied the first three of these types of evaluations to BARS development. These types of evaluations are applicable to the user interface of essentially any interactive software application.

2.1. Expert usability evaluation

The process of identifying potential usability problems by comparing a user interface design to established usability design guidelines is called *expert usability evaluation* (or *heuristic evaluation* or *usability inspection*). Those identified problems are then used to derive recommendations for improving that design. This method is used by usability experts to identify critical usability issues early in the development cycle, so that these design issues can be addressed as part of the iterative design process [11]. Often the usability experts rely explicitly and solely on established usability design guidelines to determine whether a user interface design effectively and efficiently supports user task performance (i.e., has high usability).

Usability experts may also rely more implicitly on design guidelines while they work through user task scenarios (typically created during domain analysis, another usability engineering activity in Figure 1) during their evaluation. Each evaluator first inspects the design alone, independently of other evaluators' findings. All evaluators then combine their data to analyze both common and conflicting usability findings. Nielsen [11] recommends that three to five evaluators performing an expert evaluation will find a majority of the most severe usability problems. He has also shown empirically that fewer evaluators generally identify only a small subset of problems and that more evaluators produce diminishing results at higher costs. Results from an expert evaluation should not only identify problematic user interface components and interaction techniques, but should also indicate *why* a particular component or technique is problematic. Results of this type of evaluation typically are not applicable across a variety of different application, since the purpose of the evaluation is to assess specific components or techniques for a specific application. This is arguably the most cost-effective type of usability evaluation, because it does not involve users.

2.2. User-based statistical evaluation

The process of performing relatively small and quick empirical studies to determine what specific design factors are most likely to affect user task performance we call *user-based statistical evaluation.* This can be especially effective when designing a user interface to support new and novel hardware, domains, and user tasks. Such evaluations typically focus on lower-level cognitive or perceptual tasks, where the importance of these tasks would be suggested by earlier activities in the usability engineering process. These studies are usually targeted at a specific part (e.g., a component or feature) of a user interface design, as opposed to the user interface as a whole. They may involve tasks that are atomic components of higher-level representative user tasks, and the tasks are often generic rather than application-specific. These evaluations are very similar to traditional human factors experiments and are guided by a well-crafted experimental design to assess user performance by varying design factors. Users perform tasks that are narrowly focused and carefully designed to study a specific user interface component or feature.

Such evaluations help refine various user interface components or features, in preparation for more comprehensive and application-specific formative evaluations. Our experiences indicate that the components designed and refined through quick and iterative user-based statistical evaluations produce mature user interface components and features that are well-suited to support overall application tasks and user task flow. Results of this type of evaluation typically are not applicable across a variety of different applications, since the purpose of the evaluation is to refine components or features for a specific application.

2.3. Formative usability evaluation

The process of assessing, refining, and improving a user interface design by having representative users perform task-based scenarios, observing their performance, and collecting data to empirically identify usability problems [6] is called *formative usability evaluation*. This observational evaluation method can ensure usability of interactive systems by including users early and continually throughout user interface development. The method relies heavily on usage context (e.g., user tasks, user environment, user profiles), as well as a solid understanding of human-computer interaction. The term *formative evaluation* was coined by Scriven [13] to define a type of evaluation that is applied during evolving or formative stages of design. Scriven used this in the educational domain for instructional design. Williges [16] and Hix and Hartson [6] extended and refined the concept of formative evaluation for the human-computer interaction and usability engineering domain.

A typical cycle of formative evaluation begins with creation of user scenarios based on domain analysis activities. These scenarios are specifically designed to explore and evaluate user tasks, information, and work flows. Representative users perform these tasks as evaluators collect both qualitative and quantitative data. *Qualitative data* include *critical incidents* [3], a user event that has a significant impact, either positive or negative, on users' task performance and/or satisfaction. *Quantitative data* include metrics such as how long it takes a user to perform a specific task, the number of errors a user makes during task performance, measures of user satisfaction, and so on. Collected quantitative data are then compared to appropriate baseline metrics, sometimes redefining or altering evaluators' perceptions of what should be considered baseline. Both qualitative and quantitative data are equally important since they each provide unique insight into a user interface design's strengths and weaknesses. Finally, evaluators analyze these data to identify user interface components or features that both support and detract from user task performance, and to suggest and prioritize user interface design changes. As with the first two types of evaluations, results of this type of evaluation typically are not applicable across a variety of different applications, since formative evaluation is designed to assess a specific application.

2.4. Summative usability evaluation

The process of statistically comparing several different systems or candidate designs, for example, to determine which one is "better," where better is defined in advance, is called *summative evaluation*. In contrast to formative evaluation, it is typically performed after a product or some part of its design is more or less complete. In practice, summative evaluation can take many forms. The most common are the comparative field trial, and more recently, the expert review [14]. While both the field trial and expert review methods are well-suited for design assessment, they typically involve assessment of single prototypes or field-delivered designs. The term *summative evaluation* was also coined by Scriven [13] for use in the instructional design field. As with formative evaluation, human-computer interaction experts (e.g., [16]) and usability engineers have applied the theory and practice of summative evaluation to interaction design with very successful results.

Our experiences have found that the empirical comparative approach employing representative users, instantiated in the summative evaluation process, is very effective for analyzing strengths and weaknesses of various well-formed, candidate designs set within appropriate user scenarios. However, it is the most costly type of evaluation because it needs large numbers of users to achieve statistical validity and reliability, and because data analysis can be complex and challenging. Unlike the other three types of evaluation we present, results of this type of evaluation typically *are* applicable across a variety of different applications, since they give comparative outcomes for different kinds of user interface components, features, and/or interaction techniques spanning a number of diverse user interfaces.

3. Development of the Battlefield Augmented Reality System (BARS)

3.1. Overview of BARS

Urban terrain is one of the most important and challenging environments for current and future peacekeepers and warfighters. Because of the increased concentration of military operations in urban areas, many future police and military operations will occur in cities. However, urban terrain is also one of the most demanding environments, with complicated three-dimensional infrastructure potentially harboring many types of risks [2].

A team of researchers from the Naval Research Laboratory and Virginia Tech are developing the Battlefield Augmented Reality System (BARS) [5, 8, 10] to mitigate these warfighting difficulties through the use of outdoor, mobile augmented reality. *Augmented reality* is a display paradigm that mixes computer-generated graphics with a user's view of the real world. An example is shown Figure 2. The user wears a see-through head-mounted display that the system tracks in six-degree-of-freedom space (position and orientation). Computer graphics and/or text are created and aligned from the user's perspective with the objects to be augmented. By providing direct, heads-up access to information correlated with a user's view of the real world, mobile augmented reality has the potential to recast the way information is presented to and accessed by a user.

A user wearing BARS is shown in Figure 3. Note the head-mounted display, which is where a user sees the augmented graphics view (such as in Figure 2), dynamically changing as the user moves around.

Figure 2. An example of augmented reality (AR), where graphical information overlays a user's view of the real world. A compass shows which direction the user is facing, the triangles indicate a path the user is following, a hidden chemical hazard is annotated, and the name of the street is given. Graphics are registered with the real world, so, for example, triangles appear to be painted onto the road surface. The result is an integrated display that allows heads-up viewing of the augmenting graphical information.

Figure 3. User wearing BARS equipment.

Mobile augmented reality has many research challenges related to the design of the user interface, one of which is the "Superman X-ray vision problem" [12], illustrated later in Figure 5. This problem encapsulates the fundamental advantages and disadvantages of mobile augmented reality. With such a system, a user has "X-ray" vision and can "see" non-visible objects (e.g., far-field objects that are occluded by near-field objects) and information about them. We have determined that this is a core scientific issue in AR (at least for urban military settings, our current application context), and

are studying how best to present these non-visible, occluded objects to the user. This very challenging problem in AR user interface design occurs because the occlusion cues must be artificially created with graphics in order to support natural human depth perception. Perceiving both relative and absolute depth is a critical task in military (and many other) situations, for a user to quickly identify and correctly perceive an object's or several objects' positions. For example, a dismounted warrior might want to know whether an friendly tank or squad is located between two specific buildings that the warrior cannot see but is currently targeting for fire (i.e., they are behind buildings the warrior can see).

3.2. BARS usability engineering plan

Figure 4 shows our plan for usability engineering activities for BARS user interface development, and indicates how all activities are interrelated. Specifically, results from one activity inform the subsequent activity. This plan is an instantiation of activities from Figure 1, addressing both design and evaluation of the BARS user interface. It allows us to iteratively improve the BARS user interface by a combination of techniques. This approach is based on sequentially performing a domain analysis, then an expert evaluation, followed by user-based statistical and formative evaluations, with iteration as appropriate within and among each type of evaluation. This plan leverages the results of each individual method by systematically defining and refining the BARS user interface in a cost-effective progression.

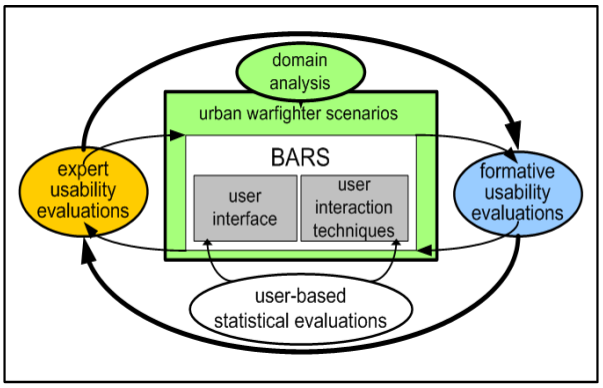

Figure 4. BARS usability engineering plan.

4. Usability evaluation activities for BARS

Team members participating in usability engineering activities for BARS include personnel from the Naval Research Laboratory (software and system developers and user interface design experts), Virginia Tech (usability engineers), Columbia University (AR user interface design expert), and a USMCR Captain, who served the critical role of subject matter expert. During usability engineering activities prior to evaluation, such as domain analysis, we created a specific scenario for BARS, to represent a realistic and significant warfighting task situation in an urban setting [5]. Then we analyzed the scenario to produce user-centered requirements.

Interestingly, producing the user-centered requirements drove an important design decision. We realized that our user-centered requirements identified a list of features that could not be easily delivered by any current AR system. For example, one BARS user-centered requirement said that the system must be able to display the location of hidden and occluded objects (e.g., personnel or vehicles located somewhere behind a visible building). This raised numerous user interface design questions related to occluded objects and how they should be presented graphically to a user (the 'X-ray vision' problem mentioned in Section 3). To address such issues, we began expert evaluations on an evolving BARS user interface design.

4.1. BARS expert usability evaluation

During six cycles of expert evaluation over a two month period, summarized in Table 1, we designed approximately 100 mockups depicting various potential designs for representing occlusion, systematically varying drawing parameters such as:

- *Lines*: intensity, style, thickness
- *Shading*: intensity, style, fill, transparency
- *Hybrid* techniques employing combinations of lines and shadings

We were specifically examining several aspects of occlusion, including how best to visually represent occluded information and objects, the number of discriminable levels (layers) of occlusion, and variations on the drawing parameters listed previously. In each cycle of expert evaluation, team members individually examined a set of occlusion representations (set size ranged from 5 to 30 mockups in a cycle), which were created using Adobe Photoshop and Microsoft PowerPoint employing video to capture real-world scenes as background images. Team members each independently performed an expert evaluation of electronically shared mockups in advance of extensive teleconference calls. During the calls, we shared our individual expert evaluation results, compiled our assessments, and collaboratively determined how to design the next set of mockup representations, informed by results of the current cycle. Because the mockups supported a very quick turn-around, we were able to evaluate many

Table 1. Summary of expert evaluations to evolve BARS user interface designs for occlusion.

Cycle No.	Purpose of this Evaluation Cycle	Medium for this Evaluation Cycle	Results / Findings
1	Initial expert evaluation and overview of BARS	BARS system	• Focus usability engineering efforts on 1. tracking and registration, 2. occlusion, and, 3. distance estimation.
2	First cut at representing occlusion in MOUT (military operations in urban terrain)	5 interface mockups including line-based building outlines and personnel representations	• Tracking study will require time to build cage (see Figure 6); focus on occlusion in the interim.
3	Examine large set of mockups that redundantly encode occlusion using various line drawing attributes	25 interface mockups systematically varying different types of line width, intensity, and style	• Line intensity and thickness appear to be the most powerful (consistently recognizable) encoding mechanisms, followed by line style. • Color and intensity of the scene can create misleading cues when using color and intensity together as an encoding scheme.
4	Continue to examine previous set of occlusion representations	25 interface mockups systematically varying different types of line width, intensity, and style	• Number of occluded layers that can be discriminably (effectively) represented by line-based encoding is three or four.
5	Examine additional visual cues to aid in distance estimation; examine use of filled polygons to represent occlusion in interior spaces	14 interface mockups using various shadings of occluded objects to show distance as well as occlusion in interior spaces	• Distance cues should be overlaid onto the ground and should be easily turned off and on by the user. • Motion parallax may help resolve some problems. • Number of occluded layers that can be discriminably (effectively) represented by shading-based encoding is three or four.
6	Examine shaded polygonal representations in a complex outdoor environment (Columbia campus), as well as hybrid designs employing lines; examine effects of motion parallax on encodings	30 interface mockups (5 mockups per set, 6 sets) systematically varying representations of occlusions employing filled (shaded) polygons, transparency, and lines. Mockups also simulated motion parallax by paging between images in a set.	• A combination of shaded polygons and line width is the most powerful encoding. • Distance encoding may be more powerful than simple occlusion. • Users should be able to push and pull the three to four levels of representation into and out of their real-world scene.

more designs than could have been implemented "live" in BARS. In fact, this use of mockups was extremely cost-effective, allowing the team to begin substantive usability evaluation work even before many BARS features were implemented.

Cycle 1 (see Table 1) served to indicate that, in fact, the mockups were an effective way of performing expert evaluations. In cycles 2 through 4, we specifically studied line-based encodings, and our results showed that line intensity appeared to be the most powerful (i.e., consistently recognizable) line-only drawing parameter, followed by line style. Further, line-based representations were discriminable at only three or four levels of occlusion. Interestingly, we found a few instances when color and intensity created misleading cues when used in combination as the encoding scheme. In cycle 5, we studied distance estimation and shading-based representations. Results indi-

cated that shading alone may not be enough to indicate distances; user-controllable overlaying of distance cues onto the ground may be necessary. Again we found that shading-based representations were also discriminable at only three of four levels of occlusion. In cycle 6, we combined both line- and shading-based representations into some hybrid designs, hoping to maximize the best characteristics of each type of representation. In particular, we found that a hybrid of shaded regions and line width, both with varying intensity, appeared to be the most powerful, discriminable representation for representing occluded objects.

Further, at this point, based on the relatively small changes we were making to the mockups, we felt we had iterated to an optimal set of representations for occlusion, so we chose to move on to formative evaluations using them. However, in retrospect (and as part of continually evolving and improving our cost-effective progression of usability evaluation – see Section 5), we realized it would have been scientifically advantageous to have run the user-based statistical evaluations next, to evolve empirically-derived user interface designs for our BARS formative evaluation. So, even though we did not perform them until after formative evaluations on BARS, we will discuss the user-based statistical evaluations next.

4.2. BARS user-based statistical evaluation

Our prior evaluations of BARS led us logically to critical design factors, in this case graphical techniques for displaying the ordering and distance of occluded objects, that needed statistical, empirical confirmation with users. Specifically, we determined from our results that a critical yet tenable set of factors and their values for a user-based statistical evaluation were:

- *Drawing style* – line, filled, line+fill (shading)
- *Opacity* – constant, increasing with levels of occlusion
- *Intensity* – constant, decreasing with levels of occlusion
- *Ground plane* – on, off

Our reasoning behind choices for each factor is detailed in [9]. The study was run with eight subjects, who saw a small virtual world that consisted of representations of three blue buildings and a red target object, overlaid, of course, on the real world. A display from one of the evaluation trials is shown in Figure 5.

The user's task was to indicate the location of the target (near, middle, or far position) as it moved among buildings from trial to trial. We examined time to perform tasks as well as task accuracy under various experimental conditions. Our results from this evaluation are reported in full in [9]. To briefly summarize, subjects made 79% correct choices and 21% erroneous choices of the target location during trials. User errors fell into two categories: the target could be closer than the user's answer, or farther than the user's answer. Subjects were most accurate when the target was in the far position; only 17.3% of their erroneous choices were made when the target was in the far position, as compared to 38.6% in the close position, and 44.2% in the middle position. Other findings indicate that the 'line+fill' drawing style yielded the best accuracy, confirming our expert evaluation results.

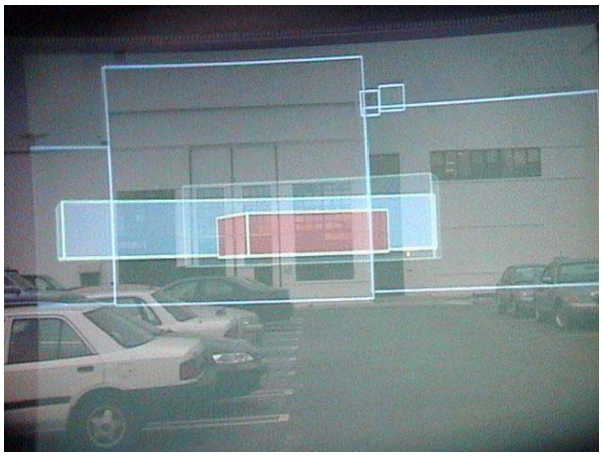

Figure 5. An example of a BARS user's view of real-world buildings augmented with overlaid graphics to indicate occluded (hidden) buildings. The overlaid information can contain text, bitmaps, or any computer-generated visual data. In this example, the lighter the shading of the object, the further away it is.

Overall, our results indicate that we evolved an effective and efficient set of graphical representations for occlusion, by applying our usability engineering methodology. These representations are being incorporated into the BARS user interface. Further, once the larger BARS user interface setting is adequately expanded and refined (i.e., using iterative expert evaluation and additional user-based statistical studies to refine other atomic user interface components and features), we expect to conduct further usability evaluations that employ comprehensive user tasks and task flows. Additional user-based statistical evaluations may, for example, study other core scientific issues in AR such as acceptable registration error (how far off the augmenting graphics can be from the real-world object) in terms of user performance and (much like occlusion) what visual representations best support distance perception and estimation for a user.

4.3. BARS formative usability evaluation

Continuing with our study of occlusion, we created a formal set of user tasks, and had five individual subjects perform the set of tasks. Three subjects were Marines and two were user interface/AR experts. The tasks were militarily relevant, inspired by our urban warfighting scenario. In the tasks, users were asked to find explicit information from the augmenting graphics that they could see. Some simple examples included answering questions such as:

- Which enemy platoon is nearest you?
- Where are restricted fire areas? Where are other friendly forces?
- Estimate the distance between the enemy squad and yourself.
- What direction is the enemy tank traveling?

Having anticipated the challenge of working in an outdoor, mobile, highly dynamic environment, team members had to consider innovative approaches to usability evaluation. Our solution was to design and build a specially-constructed motion tracking cage so that BARS could accurately track the user and accurately register graphics onto the real world. The cage provided a mounting platform for Intersense IS900 tracking rails, which are currently in common use for AR tracking. While clearly not usable in a final, fielded outdoor AR system, mounting the tracking rails on top of the cage gave us adequate tracking performance to meet our user task requirements, without waiting for completion of a totally mobile outdoor prototype AR tracking system with the required performance. The main tradeoff was that the user was not able to freely walk large distances, as envisioned in the final BARS. We therefore focused on tasks related to scanning the urban environment from the area covered by the tracking cage. Our setup also included auxiliary evaluator's monitors to provide evaluators an accurate display of a user's view. Our outdoor BARS evaluation equipment setup is shown in Figure 6.

Our overall formative evaluation results showed that users performed approximately 85% of the tasks correctly and efficiently with less than 10 minutes of training using BARS. Users liked having multiple views of various graphical augmentations, and liked being able to develop strategies to manipulate the scene and understand how BARS works. They stated that they were able to gain situation awareness from using BARS. Users disliked use of wireframes (lines) as the main augmentation representation, saying that it made the scene too cluttered. They also disliked some of the controls for manipulating augmentations (e.g., making them appear/disappear), but these controls are temporary, only for our evaluation studies, and are not intended to be included in a deployable BARS. Many of our results supported findings from our earlier expert evaluations, such as that objects must be perceived as three-dimensional and our hypothesis that no more than three or four levels of occlusion are discriminable. We made new findings such as the fact that three-dimensionality of occluded objects was easier to perceive in shaded objects than in line-drawn objects. All users had a very positive, enthusiastic reaction to BARS and its capabilities. Our experience during the formative evaluation led us to determine that the problem of representing occluded objects in AR required more attention, and specifically required us to design studies to determine what visual design factors (for occluded objects) were most effective, independent of other user interface components (e.g., text labels).

Figure 6. Outdoor tracking cage setup for BARS formative evaluation study. The cage has overhead tracking rails (barely visible under the blue canopy) so that the augmenting graphics can change as a user moves around.

4.4. BARS summative usability evaluation

We are still performing user-based statistical evaluations and formative evaluations on the BARS user interface. There is still much work to be done on the occlusion issue, as well as a variety of other challenges including tracking / registration error and distance perception / estimation. As such, we have not yet conducted comparative summative evaluations of the BARS user interface.

5. Conclusions: A cost-effective usability evaluation progression

As depicted in Figure 7, our work over the past several years has shown that progressing from expert evaluation to user-based statistical evaluation to formative evaluation to summative evaluation is an efficient and cost-effective strategy for assessing and improving a user interface design [7].

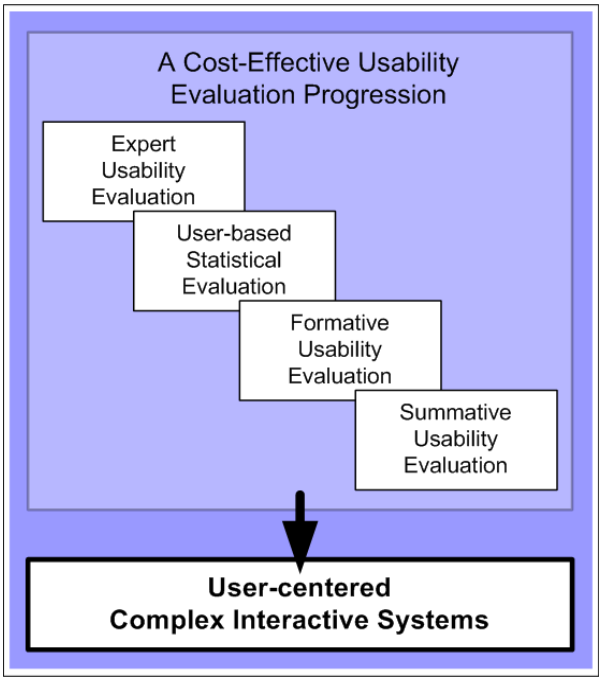

Figure 7. A cost-effective usability evaluation progression.

Our expert evaluations of BARS identified obvious usability problems or missing functionality early in the BARS development life cycle, thus allowing improvements to the user interface prior to performing user-based statistical and formative evaluations. If expert evaluations are not performed prior to user-based statistical and formative evaluations, these evaluations will typically take longer and require more users, and yet reveal many of the same usability problems that could have been discovered by less expensive expert evaluations. In cases where user interface design demands that new metaphors, interaction techniques, or user interface components be created, user-based statistical studies are an efficient method for determining what design factors are most critical for a particular user interface component or feature. These refined components can then be migrated into a mature user interface that is primed for formative usability evaluation.

Once evolving user interface designs have been expertly and formatively evaluated, then experimenters can have confidence that those designs are comparable in terms of their usability, and thus lead to a compelling comparative summative study. Otherwise, the expensive summative evaluations may be essentially comparing "good apples" to "bad oranges" [7]. Specifically, a summative study of different application interfaces may be comparing one design that is inherently better, in terms of usability, than the other ones. Developing all designs used in a summative study following our suggested progression of usability engineering activities should lead to a more valid comparison. Moreover, in our BARS work, we found that results from our user-based statistical studies are efficiently driving user interface design for our formative evaluations. We further expect the formative evaluations, in turn, to inform the design of summative studies by helping determine critical usability characteristics to evaluate and compare.

While this paper reports only on our usability engineering activities with BARS, we have been involved with and led these activities for a broad variety of applications over the past two decades (e.g., [4, 7, 9, 10]). A continual and overarching goal of all our usability engineering work is to develop, apply, and extend methods for improving the usability of interactive software applications. In particular, we have focused on developing, applying, and extending when necessary, high-impact processes for evaluating usability. Our work has produced a cost-effective progression of usability engineering activities.

6. Acknowledgments

We gratefully acknowledge research support from The Office of Naval Research and the Naval Research Laboratory and a faculty startup grant from the University of California, Santa Barbara. The BARS application has support from the Office of Naval Research under Program Manager Dr. Larry Rosenblum. Dr. Helen Gigley and Dr. Astrid Schmidt-Nielsen, also through ONR funding, have supported evolution of the usability engineering process. Others too numerous to mention individually contributed to development of these applications. Dr. Richard E. Nance, Ms. Trish Hubble, and Ms. Joyce Moser of Virginia Tech's Systems Research Center have also given much support to our efforts over the past few years.

7. References

[1] F.P. Brooks, The Mythical Man-Month: Essays on Software Engineering, 2nd Edition, Addison-Wesley, 1995.

[2] M. Dewar, War in the Streets: The Story of Urban Combat from Calais to Khafi, David & Charles, 1992.

[3] E.M. del Galdo, R.C. Williges, B.H. Williges, and D.R. Wixon, "An Evaluation of Critical Incidents for Software Documentation Design," in *Proc. Thirtieth Annual Human Factors Society Conference*, Anaheim, CA, 1986.

[4] J. L. Gabbard, D. Hix, and J.E. Swan II, "User-Centered Design and Evaluation of Virtual Environments," invited paper in *IEEE Computer Graphics and Applications*, Nov/Dec 1999 (Vol. 19, No. 6), pp. 51-59, 1999.

[5] J.L. Gabbard, J.E. Swan II, D. Hix, M. Lanzagorta, M.A. Livingston, D. Brown, S. Julier, "Usability Engineering: Domain Analysis Activities for Augmented Reality Systems," in *Proc. SPIE* Vol. 4660, p. 445-457, *Stereoscopic Displays and Virtual Reality Systems IX*, Andrew J. Woods; John O. Merritt; Stephen A. Benton; Mark T. Bolas; Eds. Photonics West 2002, Electronic Imaging Conference, San Jose, CA, 2002.

[6] D. Hix and H. R. Hartson, Developing User Interfaces: Ensuring Usability through Product & Process, John Wiley and Sons, Inc., 1993.

[7] D. Hix, J.E. Swan II, J. Gabbard, M. McGee, J. Durbin, and T. King, "User-Centered Design and Evaluation of a Real-Time Battlefield Visualization Virtual Environment," in *Proc. IEEE VR'99 Conference*, Houston, Texas, 1999. (Winner of "Best Paper" award at this conference.)

[8] S. Julier, Y. Baillot, M. Lanzagorta, D. Brown, L. Rosenblum, "BARS: Battlefield Augmented Reality System," *NATO Symposium on Information Processing Techniques for Military Systems*, Istanbul, Turkey, 2000.

[9] M. A. Livingston, J. E. Swan II, J. L. Gabbard, D. Hix, T.H. Höllerer, S.J. Julier, Y. Baillot, and D. Brown, "Resolving Multiple Occluded Layers in Augmented Reality," in *Proc. International Symposium on Mixed and Augmented Reality* (*ISMAR)*, Tokyo, Japan, 2003.

[10] M.A. Livingston, L.J. Rosenblum, S.J. Julier, D. Brown, Y. Baillot, J.E. Swan II, J.L. Gabbard, D. Hix, "An Augmented Reality System for Military Operations in Urban Terrain," in *Proc. Interservice / Industry Training, Simulation, & Education Conference (I/ITSEC '02),* Orlando, FL, 2002.

[11] J. Nielson, *Usability Engineering*, Academic Press, 1993.

[12] Personal communication, Thomas Caudell and Henry Fuchs, 1993 – present.

[13] M. Scriven, "The Methodology of Evaluation," in R. E. Stake (Ed.), *Perspectives of Curriculum Evaluation*, American Educational Research Association Monograph, Rand McNally, 1967.

[14] F. Stevens, L. Frances, and L. Sharp, User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering, and Technology Education, NSF 93-152, 1997.

[15] E.J. Swan II and J.L. Gabbard, "Survey of Augmented Reality Literature with Respect to HCI and User-Based Studies," in preparation, 2004.

[16] R.C. Williges, "Evaluating Human-Computer Software Interfaces," in *Proc. International Conference on Occupational Ergonomics,* 1984.