Available online at www.sciencedirect.com
SCIENCE DIRECT®
Interacting with Computers 16 (2004) 831–849
Interacting with Computers
www.elsevier.com/locate/intcom

Heuristic evaluation of virtual reality applications

Alistair Sutcliffe*, Brian Gault

Centre for HCI Design, School of Informatics, University of Manchester, P.O. Box 88, Manchester M60 1QD, UK

Received 25 July 2003; revised 30 April 2004; accepted 1 May 2004

Available online 20 June 2004

Abstract

This paper presents a heuristic method for evaluating virtual environment (VE) user interfaces. The method is based on Nielsen's [Usability Inspection Methods, 1994] usability heuristics, extended by VE-specific principles proposed by Sutcliffe and Kaur [Behaviour and Information Technology 19 (2000) 415–426]. Twelve heuristics are presented which address usability and presence issues. An inspection-based evaluation method is described and illustrated with three usability case study assessments, the last of which rates the applicability and validity of the heuristics by several evaluators. Use of the method uncovered several usability problems and trapped the most serious errors. Finally, VE applications integrating measures of usability and presence are discussed.

Keywords: CAVE; Virtual environment; Heuristic evaluation; Usability

1. Introduction

Several studies have highlighted usability problems associated with the use of virtual environments (VEs) (Gabbard and Hix, 1997), while field studies of VR designers have demonstrated the need for HCI knowledge and methods (Kaur et al., 1996). Others have shown that the designers of VE systems cannot rely solely on the methods developed for standard graphical user interfaces (GUIs) since their interaction styles are radically different from standard user interfaces (Bowman and Hodges, 1997; Poupyrev and Ichikawa, 1999). Most studies have followed observation and expert interpretation of users' errors (Hix et al., 1999) or experimental studies reporting performance data and problems in a range of VE technology (Bowman et al., 1999). Design principles have been

* Corresponding author. Tel.: +44-0-161-200-3315; fax: +44-0+161-200-3324.
E-mail address: a.g.sutcliffe@co.umist.ac.uk (A. Sutcliffe).

0953-5438/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2004.05.001

applied to evaluate desktop VR applications (Johnson, 1998); and checklist evaluation methods, based on Nielsen's heuristics (1994), have been adapted for VR (Kalawsky, 1999); however, no evaluation heuristics have been proposed specialised for VEs.

Quality assessment of VEs has also focused on assessment of presence, i.e. evaluating how real or natural the user's experience was when immersed in the environment. Presence has been evaluated by questionnaires which ask users to rate various qualities of the VE ranging from perceptions of 'being there' (Slater et al., 1996) to more detailed inventories ranking controls, feedback, perception of realism and user engagement (Witmer and Singer, 1999). While presence measures can benchmark VE designs in terms of their realism and overall user experience, they do not help to diagnose design flaws for formative evaluation. In this paper, we propose a set of heuristics and an expert evaluation method that follows the current widely accepted approach for user interface evaluation (Nielsen, 2000), but extend it specifically for VEs. We also attempt to unify the formative evaluation with presence assessment in VR. The paper describes the heuristic evaluation method for VR and two case studies in which it has been applied. The paper is structured in three sections. First, we describe the heuristics and the evaluation method, and then two case studies illustrating its application. This is followed by an assessment of the method by multiple experts to assess the reliability of the heuristics and the utility of the method itself. The paper concludes with a discussion of future research on evaluating VEs.

2. Heuristics for VE evaluation

Usability inspection is defined as "the generic name for a set of methods based on having evaluators inspect or examine usability-related aspects of a user interface" (Nielsen and Molich, 1990). Inspection-based methods may use guidelines or checklists as criteria to discover usability problems (e.g. ISO, 1997, part 16; Ravden and Johnson, 1989), although deciding which guidelines are applicable to particular problems becomes more difficult as the number of guidelines increase. In contrast, heuristic evaluation methods are quicker to use since they employ a limited set of design principles or heuristics. Since heuristic evaluation is quick, it is a cost-effective method and traps a high proportion of usability problems with 4–5 trained evaluators (Nielsen, 1993), although this approach is not as effective as usability testing with real users (Wharton et al., 1994). Heuristic evaluation can be performed by one usability expert although studies have shown that the effectiveness of the method is significantly improved by involving multiple evaluators (Nielsen, 1993).

The heuristics used in these studies are derived from Nielsen (1994) and our previous work on VR design principles (Sutcliffe and Kaur, 2000). Nielsen's (1993) heuristics formed the basis on which we developed the VR customised heuristics, some of which have a direct mapping, e.g. match between system and real world and our heuristics 1 and 2; others have an indirect influence, e.g. consistency and standards, user control and freedom, and visibility of system status. However, our heuristics were motivated by the different nature of VEs, in particular, the need for intuitive interaction and the sense of immersion, which is important for many VR applications that aim to simulate reality as faithfully as possible (Stone, 2002). The heuristics, with a brief

explanation, are:

1. *Natural engagement*. Interaction should approach the user's expectation of interaction in the real world as far as possible. Ideally, the user should be unaware that the reality is virtual. Interpreting this heuristic will depend on the naturalness requirement and the user's sense of presence and engagement.
2. *Compatibility with the user's task and domain*. The VE and behaviour of objects should correspond as closely as possible to the user's expectation of real world objects; their behaviour; and affordances for task action.
3. *Natural expression of action*. The representation of the self/presence in the VE should allow the user to act and explore in a natural manner and not restrict normal physical actions. This design quality may be limited by the available devices. If haptic feedback is absent, natural expression inevitably suffers.
4. *Close coordination of action and representation*. The representation of the self/presence and behaviour manifest in the VE should be faithful to the user's actions. Response time between user movement and update of the VE display should be less than 200 ms to avoid motion sickness problems.
5. *Realistic feedback*. The effect of the user's actions on virtual world objects should be immediately visible and conform to the laws of physics and the user's perceptual expectations.
6. *Faithful viewpoints*. The visual representation of the virtual world should map to the user's normal perception, and the viewpoint change by head movement should be rendered without delay.
7. *Navigation and orientation support*. The users should always be able to find where they are in the VE and return to known, preset positions. Unnatural actions such as fly-through surfaces may help but these have to be judged in a trade-off with naturalness (see heuristics 1 and 2).
8. *Clear entry and exit points*. The means of entering and exiting from a virtual world should be clearly communicated.
9. *Consistent departures*. When design compromises are used they should be consistent and clearly marked, e.g. cross-modal substitution and power actions for navigation.
10. *Support for learning*. Active objects should be cued and if necessary explain themselves to promote learning of VEs.
11. *Clear turn-taking*. Where system initiative is used it should be clearly signalled and conventions established for turn-taking.
12. *Sense of presence*. The user's perception of engagement and being in a 'real' world should be as natural as possible.

The principles of natural engagement, natural expression of action and sense of presence were motivated by questionnaire-based techniques for assessing the sense of immersion in VR environments (Witmer and Singer, 1999; Slater et al., 1996). The sense of immersion, or presence, is enhanced by a close correspondence between the VE and the user's experience of the equivalent real world. Compatibility with the user's task and domain follows the recommendations for task fit (Johnson, 1998), while heuristics 4–7

(close coordination, realistic feedback, viewpoints and navigation support) were motivated by taxonomy of VR guidelines proposed by Gabbard and Hix (1997). Heuristics 7–11 map more directly to Nielsen's heuristics for GUI interfaces, although consistent departures draws attention to the problem of using visual and audio feedback as a substitute for the sense of touch. Clear turn-taking applies to conversational VEs in which avatars may communicate with the user or when the system takes the initiative.

Underlying several of the heuristics is the assumption that the VE's role is to represent the real world as faithfully as possible. We contend that in the majority of VEs that is the case (see Stanney, 2002); however, there are VEs which represent unnatural worlds, for instance virtual molecules, or virtual information spaces. In these cases, heuristics for naturalness need to be interpreted with reference to the fit between the user's model of the task and domain, and the virtual world (heuristic 2).

2.1. Evaluation method

Our method follows Nielsen's recommendations for expert evaluation, with some differences. Evaluators familiarise themselves with the application, carry out a set of representative tasks, list problems encountered, then use the heuristics to interpret and classify the problems. Heuristic evaluation depends on the evaluator being able to establish a set of representative tasks. Since VEs are constructed to represent the real world, user tasks should ideally mirror real world actions; however, in practice limitations of technology mean that some compromises have to be accepted. Even when the VE represents an artificial world such as a complex information space, the users' ability to move in the VE will necessitate mapping real world actions to VR technology. We have therefore introduced an additional step to expert evaluation for VR, a technology audit that establishes the baseline of what the VE can reasonably be expected to deliver, given the interactive devices present in the application. The technology audit is carried out in the familiarisation period when the evaluator explores the VE and notes the presence or absence of features in the following categories, and any problems associated with them.

Operation of the user's presence. The user may be represented in the virtual world by a simple cursor or more commonly by a hand or a whole body avatar. The presence may be controlled by a variety of devices ranging from 3D mouse, space ball, joystick to pinch gloves and less frequently whole body immersion suits. The user's presence and controls can cause many problems since they provide a less than perfect rendering of the user's natural action. Suitability of the presence needs to be judged in relation to the user's task. For simple navigation, no presence may be necessary; for manipulations, however, a virtual hand is usually necessary.

Lack of haptic feedback. True virtual prototypes have no haptic feedback (sense of touch) so the user's presence can pass through representations of solid objects. To mitigate for the absence of haptic feedback, many applications use visual feedback with collision detection algorithms to prompt users when objects are selectable or have been selected. Problems caused by absence of haptic feedback may be observed with complex manipulations and physical tasks. These problems can be avoided by designing augmented reality in which interactive surfaces are modelled as physical mock-ups, but in many VEs this is too expensive.

Interactive techniques. Many VEs implement controls that allow users to fly through VEs to reach and select distant objects by ray-casting. This can be taken further by providing magic 'snap-to' effects so nearby objects automatically jump into the user's hand. These effects can cause usability problems when they are poorly designed.

Realistic graphics. VEs may not do justice to the presentation of the prototype, since most applications are not rendered in photorealistic detail. Although some evidence suggests that people can perform tasks naturally without detailed visual representations (Gabbard and Hix, 1997), graphical detail will be important for information displays and for tasks when the system environment is visually complex.

Once the technology audit has been completed, the evaluator completes a set of typical user tasks noting any difficulties encountered. These problems are then interpreted with the heuristics to assess the quality of presence (heuristics 1, 12 with contributions from 2 to 6) and diagnose design features responsible for the problems encountered (heuristics 2–11). Problems can be associated with more than one heuristic, in which case the attribution is assigned to the heuristic which explains the error most directly, followed by supplementary explanations. The following checklist guides attribution of problems to classes of design features:

- *Graphics display*, 3D depth or perspective distortion, poor resolution of image. Indicated by perceptual difficulties.
- Moving and manipulating the *user presence*, sub-divided into the hardware device being used (e.g. glove, joystick, 3D mouse, etc.) and the representation of the user in the VE. Indicated by navigation and manipulation difficulties.
- Interaction with *objects and tools* in the VE. Indicated by unsuccessful attempts to act; or poor feedback misleads users.
- *Environmental features*. Parts of the environment which created unexpected effects such as moving through walls and floating objects.
- *Interaction with other controls*, such as floating menus and palettes.
- Other *hardware problems*, such as with head-mounted display (HMD) and shutter glasses.

Once the evaluator has diagnosed the problems by assigning them to heuristics and design features as far as possible, the final stage is to rank the severity of the errors by heuristic. Indications of the severity of the identified problems are given, ranging from poor design with a severe impact likely to result in task failure to a minor problem probably curable by training. This ranking reflects the number of errors assigned to each heuristic with the evaluator's judgment about the severity of those errors, on a four-point scale:

Severe. The problem encountered would make it impossible to complete the task successfully.
Annoying. The problem would disrupt the user's task but most users would learn how to cure the error given an explanation, and some might find a work-around with time.
Distracting. The problem would disrupt the user's tasks but most users would discover the fix relatively quickly given a hint.

Inconvenient. The problem could disrupt the user's task but most users would discover the fix unaided.

The rankings provide a summative evaluation of the VE as well as a formative evaluation by prioritising areas for redesign in the next version, following the normal practice of heuristic evaluation (Nielsen, 1993).

3. Case studies

Two case studies were carried out to test the effectiveness of the VE heuristic evaluation method. The evaluator in these studies has 4 years experience in HCI research including a PhD in evaluation methodology, so he could be classified as an HCI expert. The evaluator had not used the VE heuristics before this study.

3.1. Evaluation of a scene of crime VE

The VE application used in this study was a fully immersive application using a HMD and hand-held interactive 3D mouse. It depicted a 3D environment of a scene of crime committed in garage premises and was developed by the REVEAL project in the University of Manchester (Hubbold et al., 1999). The user is able to explore the garage layout and thereby gather information about the scene of the crime. At several points within the VE application, pictures from the real environment have been inserted to help investigation of crime scenes. The representative task was to locate and enter the office in the garage premises in which the crime had been committed and note details which might be important for the prosecution, e.g. traces of blood, open window, disturbed objects, etc. The task involved the evaluator walking through the VE several times, inspecting details in the side room adjoining the garage, and finding the photographs embedded in the VE (see Fig. 1).

3.1.1. Technology audit

The VE controls were inspected during the familiarisation period when the expert explored the application to produce the following technology audit:

- *Operation of the user's presence*. The user's presence was minimal since the user's movement and viewpoint controlled all the displayed area. The exogenous viewpoint with no user presence was sufficient for navigation and inspection tasks; however, an avatar presence might be needed if the user wanted to investigate physical movements, such as exit through a small window.
- *Haptic feedback*. Since the application did not require manipulation of objects, no haptic feedback was necessary, although it would be if objects in the VE were moved and investigated.
- *Interactive techniques*. The restricted size of the VE and lack of interactive objects meant no controls or interactive facilities were necessary.

(a)

(b)

Fig. 1. Screen shot of the scene of crime VR showing (a) plan overview of the VE, and (b) the garage layout with an embedded scene of crime photograph.

- *Realistic graphics*. Since the application demanded inspection of detail, high fidelity graphics were required.

Interactive support in the VE was clearly limited, so problems needed to be interpreted in the perspective of a simple navigate and explore environment. The VE was explored thoroughly and details inspected as implied by the task. Six major problems were

Table 1
Heuristics rating and interpretation of problems encountered

Heuristic		Rating	Problems encountered
1.	Natural engagement	3	The VE maintains a fair degree of realism, but many contents appear to float. Photographs of the real environment make the VE appear less natural
2.	Compatibility with the user's task	3	Being able to fly through the air and being able to pass through solid objects, e.g. walls and the floor, not expected
3.	Natural expression of action	2	Interaction with objects to explore the scene, e.g. moving furniture, was not possible
4.	Close coordination	3	Some graphics rendering delays interfered with engagement during navigation. Also the slow constant speed of navigation was somewhat annoying
5.	Realistic feedback	N/A	No interaction with objects supported
6.	Faithful viewpoints	4	Generally good although some problems noted under heuristic 4
7.	Navigation and orientation support	3	Ability to walk through walls caused disorientation; slow pace annoying
8.	Clear entry and exit points	N/A	Not relevant for immersive VR apart from changed environments, exit from VE in desktop VR
9.	Consistent departures	3	Realistic photos incongruent
10.	Support for learning	N/A	Not necessary for simple navigation
11.	Clear turn-taking	N/A	Single user VE, no avatars
12.	Sense of presence	4	Reduced by jerkiness in display and contrast between photos and VE graphics; see heuristics 1, 4 and 7

encountered leading to the analysis of usability problems with reference to the heuristics, as shown in Table 1.

Most of the problems encountered were caused by poor navigation support and the technological limitation of delays in updating the 3D graphics display when moving and changing viewpoint in the VE. The attribution of the problems and severity rating is given in Table 2. The potential problems are also classified as requirements where clarification

Table 2
Classification of problems encountered with severity ratings and suggested design improvements

Feature	Problem description	Problem rating	Design change
Graphics	Rendering delays, floating objects	Inconvenient	Faster hardware
Presence	N/A		
Interaction	Explore objects	Distracting	Requirement clarification
Environmental features	Pass through surfaces	Annoying	Software: add movement constraints
Controls	Incongruent photos	Distracting	Software: provide ± photo controls
Hardware	N/A		

was needed from the user; software problems which could be rectified by the VE application designer; and technology problems beyond the designer's control, e.g. hardware.

The graphics rendering problems were an inconvenience which could be cured by running the application on a faster machine or improving the rendering algorithm. Floating objects were a perceptual distraction caused by poor shadowing. The interaction problem of not being able to move objects to explore the VE depended on interpretation of the user's requirements. If the VE was intended to support further investigation, then there was a case for supporting such interaction; however, this raises a scope of modelling problem: how many objects should be movable and which hidden objects should then be visible? The user specification has to indicate the extent of the real world to be modelled in the VE, and this will depend on the extent of knowledge of the crime when modelling occurs. Given this limitation, a requirement for a less interactive aide-memoire system seems to be acceptable. The disorientation problems encountered when passing through walls and other surfaces could be avoided by adding movement constraints on the user's presence, although some (audio) feedback might be advisable to signal this limitation to the user.

3.2. Evaluation of chess game VE

The second VE application was a chess game developed in a fully immersive CAVE system (Cruz-Neira et al., 1992, 1993) equipped with shutter glasses to give users stereoscopic views, and pinch gloves for manipulation. Head-tracking devices mounted on the users' shutter glasses controlled the CAVE viewpoint according to the users' body and head movements and corresponding devices tracked the users' hands.

The application displayed 12 chess pieces on a board with minimal background (see Fig. 2). The user's tasks were to move the chess pieces from a random layout to the target arrangement shown in the CAVE display. Haptic feedback for piece selection was substituted by colour changes so when a user operated the pinch glove to select a chess

Fig. 2. Selection of a chess piece.

Fig. 3. Passing a chess piece from hand to hand.

piece, the piece changed colour (from white/black to yellow) in response to the user's action (Fig. 2). The selected piece then moved in tandem with the user's hand until it was released. Once the piece was released it reverted to its original colour.

When an already selected chess piece (coloured yellow) was correctly positioned to be gripped by the other hand it changed colour from yellow to blue. The user released the piece with their first hand and the chess piece colour changed back to yellow, indicating that it continued to be selected by the second hand (Fig. 3).

The representative task was arranging the initially scattered chess pieces into an ordered arrangement with black and white pieces placed on their correct starting positions for a chess match. This was initially performed by single-handed interaction and then repeated by passing a piece from hand to hand, to evaluate more complex manipulation. The tasks were completed under two different conditions: (i) VE application with the virtual hand present; and (ii) VE application without the virtual hand being present. In both conditions, a precise bounding box provided collision detection when the virtual hand intersected with the surrounding volume of each chess piece. When the virtual hand was not represented it was necessary for the user to infer the offset between the real hand and its virtual position in the CAVE.

3.2.1. Technology audit

Operation of the VE was investigated during a learning period, resulting in the following technology audit:

- *Operation of the user's presence*. The user's presence was a virtual hand; however, it did not show any effects of pinch movements so this limited visual feedback. Only thumb and index finger pinch operations were supported.
- *Lack of haptic feedback*. Haptic feedback was substituted by colour changes in the chess pieces to indicate they were selectable. Colour changes were triggered by a collision detection algorithm when the user's virtual hand was close to the piece.

Table 3
Heuristics rating and problems encountered

Heuristic		Rating	Problems encountered
1.	Natural engagement	3	Possible to pass the chess piece through the chessboard; passing pieces between hands difficult; display distortion
2.	Compatibility with the user's task	2	Unclear colour changes when passing chess pieces from hand to hand
3.	Natural expression of action	3	Lack of haptic feedback meant colour changes and glove pinch actions had to be learned; see also heuristics 1 and 2
4.	Close coordination	4	Intermittent graphics rendering delays somewhat annoying
5.	Realistic feedback	3	Colour changes substituted for haptic feedback were not clear
6.	Faithful viewpoints	4	Generally good although could move out of the VE world
7.	Navigation and orientation support	4	Ability to walk through walls caused minor dissonance
8.	Clear entry and exit points	N/A	Step out of CAVE
9.	Consistent departures	0	No change in initiative. Colour changes for chess piece manipulation were consistent
10.	Support for learning	4	Consistent visual cues helped learning of object manipulations
11.	Clear turn-taking	N/A	Single user VE, no avatars
12.	Sense of presence	4	Reduced by some rendering delays and sparsity of overall display; see also heuristics 1 and 2

- *Interactive techniques*. The interactive objects were manipulated by pinch operation and moved by hand/arm or body movements. The position of the user hand and head/body were tracked independently by sonic devices. The small size of the VE did not require any navigation support.
- *Realistic graphics*. The application had a very sparse representation of the chess pieces and minimal environment.

The VE was clearly limited by the lack of haptic feedback for interactive support, so grip and manipulation problems needed to be interpreted in this perspective.

Carrying out the task of picking up, passing and replacing all the pieces (five pieces for each colour) produced a list of seven major usability problems, which were subject to heuristic analysis, as shown in Table 3.

Most of the problems encountered were caused by learning the rules for manipulating the chess pieces with the colour changes substituting for the lack of haptic feedback. The classification of user problems is given in Table 4.

Graphics problems occasionally caused attention to be distracted by distortion effects within corners of the CAVE, but such problems are known limitations of room- or cube-shaped CAVE environments and can only be cured by CAVE display domes. Graphics rendering delays can be cured by improved hardware or faster rendering algorithms. When the virtual hand was absent perceptual problems were much worse because the position of the virtual hand in the 3D graphic space could not be directly mapped to the user's visible real hand. Problems encountered with use of the pinch glove to manipulate objects all emanated from use of colour changes that substituted for absence of haptic feedback.

Table 4
Classification of problems encountered with severity ratings and suggested design improvements

Teature type	Description	Problem rating	Design change
Graphics	Rendering delays, display distortion	Inconvenient	Faster hardware
Presence	Grip, manipulation problems	Annoying	Software: improve colour changes
Interaction	Manipulating object	Distracting	Hardware: provide haptic feedback
Environmental features	Pass though surfaces	Annoying	Software: add movement constraints
Controls	N/A	–	–
Hardware	N/A	–	–

The colour changes needed some improvement; however, the glove pinch operations were reasonably naturally given the technological limitations. Absence of the virtual hand made interaction problems worse because the user had to learn the offset between their visible real hand and the invisible bounding box that represented their presence. Distortion in 3D depth made this learning process difficult. The ability to pass through surfaces caused some dissonance and this could be cured by adding movement constraints, although these may have to be made explicit to the user by audio feedback. Overall, this VE received a reasonably favourable evaluation given the limitations of the available technology.

4. Assessment with several evaluators

The VE application used in this study was the same as the previous case study; however, the application designer had taken the results of the first evaluation into account and changed the visual cues. When a piece was placed on the chessboard, the square on which it was placed changed colour from white/blue to dark red (see Fig. 4). When the user released the piece, the square reverted to its original colour. Also, a movement constraint was introduced to prevent chess pieces from being able to pass through the chessboard.

Seven undergraduate students (6 males, 1 female) from UMIST took part in the study. All were taking the advanced HCI course so they had knowledge of evaluation techniques and Nielsen's heuristics; however, they had only been introduced to the VR heuristics 1 week before the experiment and they had no prior knowledge or familiarity with VE applications. The evaluators were asked to complete the same task in the CAVE and follow the method using the 12 heuristics listed earlier. However, they did not test with the no-virtual hand condition and did not complete the technology audit phase of the method, although the results of the audit were presented to them to help interpretation of problems.

The task required evaluators to pick, move, pass, and replace a single chess piece on the chessboard. The initial list of problems encountered were:

- Perceptual depth difficult to judge in places.
- Problems in placement phase of chess piece manipulation.

Fig. 4. Placing a chess piece on the chessboard, after design modifications.

- Problems observed when passing chess piece from hand to hand, i.e. difficult to see colour change.
- General jerkiness and intermittent delay in graphics rendering.
- Absence of haptic feedback.

The evaluators in this study also identified problems with the application environment, where typical comments included: 'The graphics are a bit distorted', 'Graphics are blurry', and 'Slow update'. Problems were also experienced in the use of the pinch gloves, where typical comments included: 'Pick up hard', 'Lag on hands was bad', and 'Found it difficult to place piece on board'.

On the completion of the task, each evaluator rated the usability of the VE for each of the 12 heuristics. The same procedure was used as before except the four-point scale was substituted with a 1 (very poor) to 7 (very good) point scale to increase the discrimination in the evaluators' judgement. They were asked to report the reasons for their decisions and any interaction problems they had observed under the relevant heuristic. These rating scores were converted into net positive values (NPV) to reflect the range of the users' assessments. A worked example of this analysis, converting a 1–7 to a −3 to +3 scale, is given in Table 5.

Table 5
Worked example of evaluators' net positive value rating of heuristic 1 for the VE application

Rating scale	1	2	3	4	5	6	7
Conversion scale	−3	−2	−1	0	1	2	3
Rating frequency (7 evaluators)	0	1	2	4	0	0	0
Product	0	−2	−2	0	0	0	0
Total net positive value (NPV) = −4							

Table 6
NPV ratings: means and ranges (on 1–7 scale) for the 7 evaluators' scores for the VE application, where 1 = poor quality and 7 = excellent quality for that heuristic

Heuristic		NPV	Mean	Range
1.	Natural engagement	−4	3.4	2–4
2.	Compatibility with the user's task	−4	3.4	2–5
3.	Natural expression of action	3	4.4	3–5
4.	Close coordination	−2	4.3	2–7
5.	Realistic feedback	−3	4.0	1–6
6.	Faithful viewpoints	6	4.3	2–6
7.	Navigation and orientation support	2	4.3	1–7
8.	Clear entry and exit points	1	4.3	0–5
9.	Consistent departures	1	4.1	3–5
10.	Support for learning	9	5.3	3–7
11.	Clear turn-taking	3	4.8	0–6
12.	Sense of presence	−3	3.6	3–4

NPVs provide an aggregate score rating to interpret mean and ranges.

The evaluators' ratings of the VE application using the 12 heuristics is given in Table 6. These show moderate variation with most ratings tending towards poor or neutral assessments. An exception to this is the relatively high NPV rating scores for heuristic 6 (faithful viewpoints) and heuristic10 (support for learning). For heuristic 6, the evaluators' rating score (6) agreed with the experts' judgement in the previous study, although they disagreed on heuristic 10. The VE application only had minimal visual cues and in addition, as cited under heuristics 2 and 5, it was possible to pass a chess piece through the chessboard, since no movement constraint was provided.

The evaluators' mean ratings and the single expert's rating (see Table 3) are similar, but did show some improvement between the evaluations. This indicates that the evaluation and the changes made by the application designer had been successful. This was reflected in comments on one of the design features which had been changed, insertion of additional visual cues: 'Visual cues are good" and 'Use of colour makes up for lack of haptic sense'. Furthermore, the NPV rating score for heuristic 3 (natural expression of action) was positive. The addition of more evaluators did not reveal more usability problems but this may have been limited by the simple nature of the application.

The evaluators' judgement for most of the heuristics was generally consistent as the distributions of scores on all heuristics had normal distributions (i.e. non-bipolar), although the low overall number of evaluators precluded statistical testing of inter-evaluator agreement. Four of the seven evaluators gave no rating for heuristic 8 (clear entry and exit points) and three evaluators gave no rating for heuristic 11 (clear turn-taking), which they considered to be not applicable. All the evaluators reported and analysed three out of the five general problems but two evaluators did not report perceptual depth and three did not comment on the absence of haptic feedback. However, when the evaluators' judgement was compared with an independent evaluation with users on the same application (Sutcliffe et al., in prep.), the granularity of the problem description in this study was not as precise as problems discovered by observing users' errors.

Table 7
NPV ratings for the utility of the heuristics by the seven evaluators

Heuristic		NPV
1.	Natural engagement	15
2.	Compatibility with the user's task	16
3.	Natural expression of action	13
4.	Close coordination	16
5.	Realistic feedback	15
6.	Faithful viewpoints	17
7.	Navigation and orientation support	6
8.	Clear entry and exit points	−7
9.	Consistent departures	7
10.	Support for learning	13
11.	Clear turn-taking	1
12.	Sense of presence	11

− 3, not useful and +3, very useful for interpreting usability.

For instance, users noted problems in selecting the chess piece caused by the collision detection algorithm, as well as in passing pieces. In common with other studies on heuristic evaluation, more than five evaluators trapped most of the serious errors; however, the heuristics did not clearly indicate the root cause of the problems.

4.1. Evaluation of the heuristics

Finally, each evaluator was asked to rate the 12 heuristics on a 1–7 scale as to how applicable/valid they considered each heuristic to have been in their evaluation, and the reasons for their decision. The evaluators' NPV ratings given in Table 7 show a moderate variation with most ratings being favourable and therefore applicable/valid assessment. However, low NPV scores were given for heuristics 8 (clear entry and exit points) and 11 (clear turn-taking), with moderate rankings for heuristics 7 (navigation and orientation support) and 9 (consistency departures). For heuristic 7, the evaluators' rating score (6) agreed with the experts' view on the heuristic's applicability. Evaluators' comments included: 'Not necessarily appropriate, if in a simple single environment'. For heuristic 8, the evaluators' rating score (− 7) was low; however, this was explained by their comments on the heuristic's applicability. Four of the seven evaluators considered the heuristic to be 'not applicable'; typical comments included: "There wasn't really the case because we just had to put the gloves and glasses on".

Two evaluators questioned the meaning of heuristic 9. Finally, for heuristic 11, the evaluators' rating score (1) was low because of its inapplicability, three of the seven evaluators considering it to be 'not applicable'.

4.2. Lessons learned

In the initial single expert study, the heuristics proved to be easy to interpret and indicated problem areas in the design; however, heuristics 8 and 9 (entry/exit points, consistent departures) were more applicable to desktop VR and applications with

embedded GUI features. This suggested that the heuristics needed to be filtered at the technology audit stage so only relevant ones were applied. There was some overlap between the first three heuristics, which all referred to slightly different aspects of natural interaction.

In the several evaluators study, the need to filter the applicability of the heuristics was also evident. Tailoring the heuristics to different types of VE will cure the 'not applicable' problem. The heuristics need to be presented as a core set for all VEs (1–7 and 12) with the addition of heuristics 8 (clear entry and exit points) and 9 (consistent departures) for desktop VR, and heuristic 11 (clear turn-taking) for collaborative VEs or when system initiative is involved. The importance of navigation support (heuristic 7) and support for learning (heuristic 10) will depend on the complexity and size of the VE. Some clarification of heuristics 1–3 will help to separate issues of overall experience of engagement (1), compatibility with users' expectations (2) and controllability of actions (3). Several evaluators reported difficulty in interpreting heuristics. This could be ameliorated by giving examples of good and bad VEs to illustrate each heuristic; however, examples might bias users towards irrelevant details. On balance, we feel increased training is the answer to interpretation, rather than provision of limited examples. Finally, there is the question whether we need additional heuristics for problems and design issues we may have omitted. The comments of our evaluators and the problems they encountered suggest that our current set is appropriate, although experience with developing technology may require enhancement of the set in the future.

The advantages of the heuristics were that they are quick to use and provide insight into usability problems by drawing attention to high-level design concerns. This role is shared with other heuristics which have been proposed for evaluating different types of user interfaces, e.g. CSCW applications (Baker et al., 2002), ambient displays (Mankoff et al., 2003). The limitations inevitably are the trade-off between rapid use and detailed advice, which can be found in taxonomies of guidelines (Gabbard and Hix, 1997). One improvement may be to add heuristics that focus attention on the major components of most VEs, i.e. the user's presence, interactive objects, and the quality of the graphical world.

5. Discussion and conclusions

The case studies demonstrated that the 12 heuristics we developed from earlier work (Sutcliffe and Kaur, 1997) provided a useful tool that performed an efficient and meaningful usability evaluation of VE applications. The heuristics augment the criteria for evaluating desktop VR proposed by Johnson (1998): task fit, navigation support and subjective satisfaction, while providing a quicker evaluation process than the VRUSE usability feature checklist (Kalawsky, 1999). Of the 10 usability factors in VRUSE, input and output devices map to the technology audit in our method, while others (error correction, consistency, user guidance) are similar to Nielsen's UI heuristics; while only two (simulation fidelity and presence) are explicitly targeted at VR applications. Overall, in the context of the three studies, we consider the heuristic evaluation process to have been a success since usability problems were identified with only a small expenditure of

effort. The heuristics represent an important extension to expert HCI evaluation methods and address the specific issues raised by VEs, in particular, the integration of presence and usability. However, we believe it is important to judge VEs from a baseline of the available technology, hence the audit part of the method played an important role. The method we proposed combines heuristics with a technology audit that focused on particular aspects of VR technology. We argue that since VR aspires to create the perfect illusion of an interactive world, but is inevitably limited by the technology for rendering interaction, it is important to benchmark the evaluation by making the limitations of the technology explicit.

In contrast to studies on Nielsen's heuristics, which demonstrated that more experts (up to six) trapped more errors within a law of diminishing returns, we did not find increasing the number of evaluators discovered many more errors. However, this might have been caused by a ceiling effect with a simple application, i.e. there were no more problems to find. As our evaluators were arguably novices, it is reassuring that they found the same errors as the single expert. This demonstrates that the method can be used by evaluators who have limited HCI training, a finding which agrees with studies on UI heuristics (Nielsen and Molich, 1990). The small number of errors in the VEs we studied may have disguised the effort of adding more evaluators to trap additional errors. Furthermore, many of the errors had the same root cause, i.e. problems in selecting and placing pieces could be traced to depth perception and the collision detection algorithm which signalled when a piece was selectable. Problems can be reported at different levels of abstraction and this can limit the effectiveness of heuristic analysis.

The approach we have adopted is similar to other extensions to Nielsen's method for CSCW (Baker et al., 2002) and ambient displays (Mankoff et al., 2003). These authors also modified Nielsen's basic set to include more abstract qualities of the design, for instance 'sufficient information design' for ambient displays (Mankoff et al., 2003) and 'provide consequential communication of an individual's embodiment' in CSCW (Baker et al., 2002). Mankoff et al.'s study demonstrated that their specialised heuristics were superior to Nielsen's in analysing errors. While we do not have comparative data, we expect our heuristics to be superior to Nielsen's for VEs; however, we believe that a combination of standard Nielsen and specialised heuristics may produce the optimal result. The heuristic method presented in this paper augmented Nielsen's approach by adding the technology audit which we argue is an important way of calibrating judgement for different technologies. The method is a generic tool for the evaluation of VE applications; however, it needs to be tailored to different styles of VE. For instance, desktop VEs frequently employed part of a GUI interface, so heuristic 8 (clear exit and entry points) is relevant, whereas for immersive VE it is not. Likewise, clear turn-taking only applies when other actors are present in the VE. In our revised method, we have included a guide to filtering the heuristics, so only appropriate questions are asked. We expect our heuristics to be one in a battery of evaluation techniques, which can be employed according to different resource constraints; for example, cooperative evaluation by observing users' problems (Monk et al., 1993). Observation of users' problems and interpretation of error causes has been demonstrated for immersive applications (Hix et al., 1999). Questionnaires will continue to play a role in summative evaluation of presence (Witmer and Singer, 1999; Slater et al., 1995), and expanding the heuristics we proposed could be used for summative

evaluation with these techniques. Finally, cognitive walkthrough approaches have also been demonstrated for VR using an extension of Norman's model of action for VR (Kaur et al., 1999). Bowman's framework for comparing evaluation methods for VR applications (Bowman et al., 2002) provides a means of locating the contribution of our heuristic method in a wider perspective of other methods, although further studies will be necessary to develop a means of selecting the optimal approach to adopt given a set of evaluation needs and resource constraints.

Acknowledgements

We would like to thank Adrian West, Toby Howard and Roger Hubbold of the Advanced Interfaces Group, Department of Computer Science, University of Manchester for the use of the application used in the first case study. We would also like to thank Terrence Fernando and Kevin Tan of the Centre for Virtual Environments, University of Salford, for their assistance in the latter two case studies, and students from the Department of Computation, UMIST for their participation. The research was supported by EPSRC grant GRM68749 Immersive Scenario-based Requirements Engineering.

References

Baker, K., Greenberg, S., Gutwin, C., 2002. Empirical development of a heuristic evaluation methodology for shared workspace groupware, In: Proceeedings of CSCW'02 New Orleans, ACM press, USA, pp. 96–105.

Bowman, D.A., Hodges, L.F., 1997. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments, Proceedings: 1997 Symposium on Interactive 3D.

Bowman, D.A., Johnson, D.B., Hodges, L.F., 1999. Testbed evaluation of virtual environment interaction techniques, Proceedings: ACM Symposium on Virtual Reality Software and Technology, London 20–22 December 1999, ACM Press, New York, pp. 26–33.

Bowman, D., Gabbard, J., Hix, D., 2002. A survey of usability evaluation in virtual environments: trade classification and comparison of methods. Presence: Teleoperators and Virtual Environments 11(4), 404–424.

Cruz-Neira, C., Sandin, D.J., De Fanti, T.A., Kenyon, R.V., Hart, J.C., 1992. The CAVE: Audio visual experience automatic virtual environment. Communications of the ACM 356, 64–72.

Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., 1993. Surround-screen projection-based virtual reality: the design and implementation of the CAVE. Computer graphics, Proceedings: SIGGRAPH93, Computer Graphics, ACM Press, New York, pp. 135–142.

Gabbard, J.L., Hix, D., 1997. A Taxonomy of Usability Characteristics in Virtual Environments, Deliverable to Office of Naval Research, grant no. N00014-96-1-0385, Department of Computer Science, Virginia Polytechnic Institute, Blacksburg, VA.

Hix, D., Swan, J.E., Gabbard, J.L., McGee, M., Durbin, J., King, T., 1999. User-centered design and evaluation of a real-time battlefield visualization virtual environment. In: Rosenblum, L., Astheimer, P., Teichmann, D. (Eds.), Proceedings: IEEE Virtual Reality '99, Houston TX 13-17 March 1999, IEEE Computer Society Press, Los Alamitos, CA, pp. 96–103.

Hubbold, R., Cook, J., Keates, M., Gibson, S., Howard, T., Murta, A., West, A., Pettifer, S., 1999. GNU/MAVERIK: a micro-kernel for large-scale virtual environments, Proceedings: VRST'99, ACM Symposium on Virtual Reality Software and Technology, ACM Press, New York.

ISO, 1997. ISO 9241: Ergonomic Requirements for Office Systems with Visual Display Terminals (VDTs), International Standards Organisation.

Johnson, C., 1998. On the problems of validating desktop VR. People and computers XIII, Proceedings: HCI'98, Sheffield 1–4 September 1998, Springer, Berlin, pp. 327–338.

Kalawsky, R.S., 1999. VRUSE: a computerised diagnositc tool for usability evaluation of virtual/synthetic environment systems. Applied Ergonomics 30(1), 11–25.

Kaur, K., Maiden, N.A.M., Sutcliffe, A.G., 1996. Design practice and usability problems with virtual environments, Proceedings: Virtual Reality World 96, Stuttgart, [Informal proceedings].

Kaur, K., Sutcliffe, A.G., Maiden, N.A.M., 1999. Towards a better understanding of usability problems with virtual environments. In: Sasse, A., Johnson, C. (Eds.), Proceedings of INTERACT 99: IFIP TC.13 Conference on Human Computer Interaction, IFIP/IOS Press, Amsterdam, pp. 527–535.

Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., Ames, M., 2003. Proceedings of Human Factors in Computing Systems CHI 2003. CHI letters 5(1), 169–176. Heuristic evaluation of ambient displays. In: Bellotti, V., Erickson, T., Cockton, G., (Eds.), CHI 2003 Conference Proceedings: Conference on Human Factors in Computing Systems, Fort Lauderdale FL 5-10 April 2003. ACM Press, New York, (CHI Letters 5(1) pp. 169-176).

Monk, A., Wright, P., Haber, J., Davenport, L., 1993. Improving your human-computer interface: a practical technique, Prentice Hall, London.

Nielsen, J., 1993. Usability Engineering, Academic Press, New York.

Nielsen, J., 1994. Heuristic evaluation. In: Nielsen, J., Mack, R.L. (Eds.), Usability Inspection Methods, Wiley, New York.

Nielsen, J., 2000. Designing Web Usability: The practice of Simplicity, New Riders.

Nielsen, J., Molich, R., 1990. Heuristic Evaluation of User Interfaces, SIGCHI Bulletin, April: Special Issue, pp. 249–256.

Poupyrev, I., Ichikawa, T., 1999. Manipulating objects in virtual worlds: categorization and empirical evaluation of interaction techniques. Journal of Visual Languages and Computing 10(1), 19–35.

Ravden, S., Johnson, G., 1989. Evaluating Usability of Human-Computer Interfaces, Ellis Horwood, New York.

Slater, M., Usoh, M., Steed, A., 1995. Taking steps: the influence of a walking technique on presence in virtual reality. ACM Transactions on Computer-Human Interaction 2(3), 201–219.

Slater, M., Linakis, V., Usoh, M., Kooper, R., 1996. Immersion, presence and performance in virtual environments: an experiment using tri-dimensional chess, Available online at http://www.cs.ucl.ac.uk/staff/m.slater/Papers/Chess/index.html.

Stanney, K., 2002. In: Jacko, J.A., Sears, A. (Eds.), The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Lawrence Erlbaum Associates, Mahwah, NJ.

Stone, R.J., 2002. Applications of virtual environments: an overview. In: Stanney, K., (Ed.), Handbook of Virtual Environments: Design, Implementation and Applications, Lawrence Erlbaum Associates, Mahwah NJ, pp. 827–856.

Sutcliffe, A.G., Kaur, K., 1997. Modelling Interaction For Virtual Reality Systems, Poster Paper: INTERACT'97, Sydney, IOS Press, Amsterdam.

Sutcliffe, A.G., Kaur, K.D., 2000. Evaluating the usability of virtual reality user interfaces. Behaviour and Information Technology 19(6), 415–426.

Wharton, C., Reiman, J., Lewis, C., Polson, P., 1994. The cognitive walkthrough method: a practitioner's guide. In: Nielsen, J., Mack, R.L. (Eds.), Usability Inspection Methods, Wiley, New York, pp. 105–140.

Witmer, B.G., Singer, M.J., 1999. Measuring presence in virtual environments: a presence questionnaire. Presence 7, 225–240.