

Usability Evaluation Techniques for Interactive Television

*Lyn Pemberton
Richard N. Griffiths*

School of CMIS
University of Brighton
Brighton BN2 9XL
East Sussex, UK
LP22@brighton.ac.uk

Abstract

Observational techniques are now well established as tools for evaluating the usability and acceptability of desktop systems such as interactive applications and web sites, whether in the user's own setting or in a usability lab. When interactive elements such as games, electronic programme guides and voting mechanisms began to be integrated into broadcast television programmes and the need arose for tools to gauge the usability of these elements, it was natural to turn to observational techniques to supplement the survey-based techniques more familiar to researchers from the broadcast TV world. In this paper, we explore how the observational techniques familiar to HCI practitioners can be used to study viewers' interactions with interactive and enhanced television. We describe the setting up of a usability lab for domestic products such as interactive television and consider the implications of this new application area for the conduct of usability studies, the analysis of the data gathered and the nature of usability in this changed context.

1. Introduction

Governments and businesses are interested in using the penetration of the TV set into Western homes to sell products and services, disseminate information, enable learning and so on via interactive TV services and facilities. "Interactive" in the television context is understood to include video on demand, t-commerce, interactive programme guides and enhanced broadcast television. However, if they are to be taken up, interactive facilities need to be usable by viewers. This brings new challenges for television programme producers, who do not have a strong tradition of fine grained analysis of viewer interaction with television, preferring instead to rely on survey methods such as diaries, questionnaires, focus groups or automated monitoring to discover viewers' attitudes (Gauntlett & Hill, 1999). Specialists in user interface evaluation, on the other hand, are well acquainted with the issues involved in testing a range of applications designed for desktop PC's, but are now faced with the prospect of modifying their techniques and methods to cope with the demands of new platforms, including interactive television. Several evaluation techniques may be applicable to iTV, including analytical approaches such as heuristic evaluation (Nielsen, 1993). However it is not certain that general desktop-oriented heuristics such as "speak the user's language", "be consistent" and so on still apply in the same way and to the same extent. Building on the growing evidence from studies reported in the literature it is beginning to be possible to derive TV-specific heuristics (Daly-Jones, 2002). However, nothing is very strongly established yet. Other analytical techniques, such as Cognitive Walkthrough (Newman & Lamming, 1995) may also be useful and are probably a very effective technique to use during the

design and development process, particularly of task-oriented elements. However, here we concentrate on empirical evaluation, based on observation and interview sessions with viewers.

2. Why interactive television is different

There are a number of areas that distinguish the use of personal computers from the use of iTV. These differences suggest that evaluating iTV might need a different approach from desktop applications. They also suggest that results that are reliable for desktop applications may need handling with more caution in an iTV context.

Physical characteristics of interaction: Viewers watch television at some distance from the screen, typically in an environment that is oriented toward relaxation and comfort. Resolution is much lower than computer display screens and colour behaves differently. Detailed information is presented via audio. All interactions are carried out via a handset (combined in some cases with a keyboard). A lab set-up is needed in which views of the handset, screen and viewer's focus of attention can be captured together with any soundtrack.

Multiple information channels are mediated via the same device: There is conflict between watching the broadcast stream and manipulation of any interactive components – viewers must divide their cognitive resources between watching and interacting, and this may be reflected in the design through allocation of screen 'real-estate'. Although an interaction element may be usable in a lab setting when the broadcast stream is not present, it may not work so well which when the viewer's attention is engaged on the broadcast stream.

Embedded nature of services: Services tend not to be stand-alone but embedded within a programme or at least a channel. The perceived usability of a service and aesthetic preference for the embedding programme correlate. The services may in principle be very usable, but if a viewer would never think of watching the programme they enhance, this finding becomes irrelevant. Interaction complexity that may be acceptable to a committed fan may deter the casual viewer.

Broadcast-related aspects: While some services such as EPG's and video on demand can be evaluated fairly easily because they are not particularly time-dependent, enhanced broadcast TV is very difficult to test before broadcast. Motivation for using the interactivity may depend on simultaneous interaction with other viewers, e.g. a tendency to interact in voting application may be heightened if viewers are told a large number of other viewers are sending in votes. The communal experience of watching with potentially millions of others will not be available in non-real-time viewing. Even for summative evaluation of a broadcast product there are problems. The programme may go out during evenings or weekends when usability labs are not normally accessible. Moreover, only one trial per broadcast can be carried out in a single lab.

Time-related aspects: The motivation to carry out a particular interaction may be determined by an arbitrary element of the broadcast. For instance, while watching a football match, a change to a preferred type of view may be required during particular aspects of play such as penalty shootouts. There may be very tight time constraints on initiating and receiving this change. The continuity of monitoring the event being presented must not be lost and at any time the change may be immediately cancelled or modified. Thus there are characteristics in common with real-time control systems. Exercising this interactivity should be at the spontaneous whim of the viewer. Instructing the test subject to use this option may mean that significant aspects of it are missed.

The optional status of television viewing: Television tends to mean leisure and entertainment rather than work or other serious pursuit. Almost by definition it is never going to be mandatory for anyone to use interactive services in the way that using a work-oriented application such as database or even a word processor might be (though exceptions might exist, e.g. for educational TV). Thus the task-oriented approach most often adopted by usability evaluators may be inappropriate. Even when a task, such as booking a holiday or checking a viewing time, is the

focus, viewers always have a choice of accomplishing it via an alternative means to the television, e.g. by using the telephone or consulting a newspaper. This means that assumptions about subjects' willingness to persevere with a task in non-controlled circumstances need to be very conservative.

Social characteristics of the interaction: The domestic setting in which TV is consumed is complex and is difficult to emulate in all its facets. People have a tendency to do other things, such as ironing, knitting or reading while viewing (Gauntlett & Hill, 1999). They often view in company (Masthoff, 2002). They may be subject to interruptions of varying frequency and significance. Use of TV interactivity may compete with other domestic resources, e.g., for digital satellite in the UK, the telephone line. Users may spend a long time using a back channel in the lab that they'd be unlikely to do at home, due to cost or pressure from other family members.

The economics of viewing: Many TV interactions have a billable cost associated with them (often the reason for their existence). Who is paying the bill is a big issue that will impact on decision as to if and when to interact and may influence the options exercised. Simulating this in the lab is extremely difficult.

3. What artefacts can we evaluate?

As in any design project, we need some sort of artefact, design or prototype before evaluation is possible. With broadcast iTV programmes, however, this is not possible. There is only one shot, which means that any evaluation must be done at broadcast time and necessarily be summative in nature. This is much less flexible even than conventional TV where a video can be used to time-shift. Interactive elements cannot be recorded and played back in this way¹.

If we are not to evaluate iTV at broadcast time, the interactivity must be simulated using some kind of **prototype**. There is some scope for using paper prototyping, particularly for the task-related elements of a service, e.g. t-commerce transactions, an EPG search or an account-related task such as changing a PIN number. The televisual aspects of the product would be lost, but this is often the case anyway in paper prototyping, where high fidelity is not the goal (Rettig, 1994). 'In principle' usability for the task can be established. Other options are prototypes in HTML, an authoring package such as Director or the target middleware for the application. All these options are relatively labour intensive, particularly some varieties of middleware. In addition there is a danger of the prototype working much more rapidly than the delivered service (due to absence of transmission delays) so it may be necessary to integrate pauses to simulate realistic performance.

4. Selection of participants

Once a prototype is available, participants need to be recruited. Participant selection may be quite complex. For many desktop Windows™ based applications, it is sufficient to choose testers with broadly similar occupational backgrounds. For iTV, the experience level will vary widely and often be confined to one specific system, e.g. in the UK, Freeview, Sky, NTL or Telewest. Experience of and attitudes to other technologies such as mobile phones, Internet and Teletext may influence performance. Subjects should obviously be from the target demographic group for the programme or service. In addition, ideally subjects should be selected who would normally view together. It is not trivial to compose a group of family or friends, particularly if the observation session must take place at the time the relevant programme is aired.

¹ An exception is in cases where interactive elements are made available over a longer period, as was the case with several of the BBC's interactive documentaries. However, in these cases, interactivity resembles Web on TV rather than any more dynamic interaction.

5. Facilities for observation

Clearly, a conventional office-type usability lab will not do for evaluating iTV. At the University of Brighton we have set up the Domestic Technology Usability Lab for evaluation of usability and accessibility aspects of domestic multimedia technology, particularly interactive television. The lab consists of a room, four metres by three metres, fitted out as a domestic lounge. It contains two and three-seat sofas, a coffee table, standard lamp and a wide screen domestic television. Set in a side wall is a two-way mirror enabling the occupants of the room to be observed without distraction. Close-circuit television cameras on motorised mounts within the room enable recording of the occupants. The television display may also be recorded directly. A microphone in the room picks up conversation. The video and audio feeds from the room are combined via a video multiplexer and digitiser to be stored on the hard disc of a PC. The format of the recorded video can be varied to suit the requirements of the test being carried out, individual stream, picture-in-picture and quad being supported. A simultaneous recording onto VHS tape can be made as a backup. Behavioural analysis software is available on the PC. This enables the marking up of the captured video stream so that frequency and timings of significant events may be obtained. To evaluate interactive television use, two cameras and the display are usually recorded in quad format. A camera above the TV set records a panoramic view of the room, showing the positions and interactions of the room's occupants. A second camera fitted with a motorised zoom lens, placed to one side of the subjects' position, records the manipulation of the TV remote control. The motorised pan, tilt and zoom provide a constant close up view of the remote enabling individual button pushes and other manipulations to be recorded.

6. Practical Issues

The most effective approach we have found to be co-discovery, where pairs or trios of users, ideally in family or friendship groups, are observed together. This makes "thinking aloud" a natural activity, allowing decision making and other processes to be captured. One needs to be aware, however, that the dynamics of the group may effect the interaction. In particular, control of the handset in couples and families may be an issue.

As far as the physical set-up is concerned, we try to encourage viewers to sit within camera shot, close enough together to pass over handset without stretching. This means that handset passing is not prevented. To create a relaxed atmosphere, refreshments need to be available, at no more than arm's length. These are tailored to fit the group: beer for football fans, fizzy drinks for children, tea for older participants and so on. Participants typically prefer to keep their belongings with them, so there is a need to ensure that handbags and coats do not obscure the camera's view of handset use.

Motivation is a major issue. Instructions and tasks need to be "very simple and readily achievable" (Daly-Jones, 2002, p. 3). It often becomes clear that participants are only persisting with a task because of the slightly formal situation in which they find themselves. At home, they would have simply reached for the listings magazine or given up the struggle. We find that with some viewer groups, particularly older and less confident viewers, having a facilitator in the room helps to keep the task going. We find it useful to have a communication channel from the observers to the facilitator in the room via headphones. However, the price of putting a facilitator in the room is creating an even more formal situation. Whoever the participants, it is useful to agree in advance a signal or form of words to mark the point at which they would naturally give up the task.

During the observation session, we find it useful to capture (and subsequently tag) the following behaviours:

- Use of handset, including passing from one participant to another
- Switches of attention between screen and handset
- Participants' noticing, commenting on and activating screen controls
- Interaction between participants, e.g. instructions, suggestions and evaluative comments
- Comments on the interaction relating to learning, recognition, remembering, forgetting and awareness of functionality
- Comments on previous models of interaction such as Teletext or Internet browsers. These are particularly interesting for eliciting mental models and terminology

Wherever reasonable, we interview participants directly after observation sessions. Things happen very quickly in sessions and asking questions afterwards clarifies the reasons for particular actions. However, it is important to be extremely tactful in these situations. Participants may be tired or feel that they have "failed" in some way if they do not complete tasks. It is important not to confront a participant with failure, especially if they have previously claimed familiarity with the technology. Another aspect of post-test interviews is that participants, perhaps to please the facilitator, will tend to give more positive evaluations of the session than are warranted from the success of the interactions, often claiming ease of use where things were clearly difficult.

7. Conclusions

Even more than desktop technologies, iTV is embedded in its situation of use. To gauge the acceptability of iTV we need to simulate this as far as possible, though much is lost in an artificial setting, even if it is carefully designed. To the normal infelicities of a lab setting we need to add the problem that participants are not their own home, not in their natural grouping, not interrupted, not watching their choice of programme, not paying with their own money or tying up their own phone line. Certainly lab observation can be used with success to deliver judgements on the in-principle usability of services. However it tells us little about the attractiveness, usefulness or likelihood of adoption by viewers. The market researchers' battery of survey techniques, complemented by automated behaviour capture software, may fill this gap.

References

- Gauntlett, D. & A. Hill. 1999. *TV Living: TV, Culture and Everyday Life*. London: Routledge.
- Daly-Jones, O. 2002. Navigating your TV: the Usability of Electronic Programme Guides. In *Usable iTV* 1/3, pp. 2-6.
- Masthoff, J. & R. Luckin (Eds.) 2002. *Proceedings of the workshop Future TV: Adaptive instruction in your living room* Intelligent Tutoring Systems conference, San Sebastian, 2 June.
- Masthoff, J. 2002. *Modeling a group of television viewers*. In Masthoff & Luckin, 2002.
- Newman, W. & M. Lamming. 1995. *Interactive System Design*. Addison-Wesley: London.
- Nielsen, J. 1993. *Usability Engineering*. Academic Press: Boston.
- Rettig, M. 1994. Prototyping for Tiny Fingers. *CACM*, April, 37/4.
- University of Brighton Interactive Technology Research Group. 2003. Domestic Technology Usability Laboratory Description. <http://www.cmis.brighton.ac.uk/Research/ITRG/DTULab/>