Technical Report No: 2002/02

Extending Low-Cost Remote Evaluation with Synchronous Communication

Lynne Dunkley

Lucia Rapanotti

Jon G. Hall

2002

Department of Computing
Faculty of Mathematics and Computing
The Open University
Walton Hall,
Milton Keynes
MK7 6AA
United Kingdom

http://computing.open.ac.uk

The Open University

Extending Low-Cost Remote Evaluation with Synchronous Communication

Lynne Dunckley
Department of Computing, Thames Valley University
Wellington Street, Slough, SL1 1YG,UK
Tel: *+44 1753 69 7739*
Email: *Lynne.Dunckley@tvu.ac.uk*

Lucia Rapanotti
Department of Computing, The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
Tel: *+44 1908 654125*
Email: *L.Rapanotti@open.ac.uk*
URL: http://mcs.open.ac.uk/lr38/

Jon G. Hall
Department of Computing, The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
Tel: *+44 1908 652679*
Email: *J.G.Hall@open.ac.uk*
URL: http://mcs.open.ac.uk/jgh23/

Write-along Low Cost Remote (LCR) evaluation has been proposed as a highly efficient method for remotely evaluating usability problems with prototype interfaces. In a previous study, it was noted that this efficiency was at the cost of a loss of the conversational nature of the evaluation present in think-aloud methods. In this paper, we assess this loss through a comparison with an extended LCR, which uses real-time conferencing tools to introduce synchronous communication. An experimental investigation was carried out on an interactive prototype interface with known usability problems. In this way the effectiveness of the new method using real-time conferencing tools could be assessed and recommendations for best practice set out.

Keywords: Remote usability evaluation, real-time conferencing tools, real world context

1 Introduction

Information systems are being developed to support increasingly complex tasks. Many of these systems are distributed, involving geographically remote users. The growth of network and, particularly, Web technology means users are

communicating with central systems using a wide range of machines and operating systems. Developers often have limited access to representative users for usability testing in laboratories. In these cases the cost of transporting users and developers to remote locations can be prohibitive. The network itself and the remote work setting are also intrinsic parts of the system which produce usage patterns that are difficult to reproduce in a laboratory. This problem is particularly acute for international software development, where the development may need to involve usability evaluations in different countries.

In order to take such factors into account, remote usability evaluations have been proposed in the literature. In general, two approaches to remote usability data collection methods have been used. The subjective approach can range from reports from users, user-identified 'critical incidents' to questionnaires, interviews and ethnographic techniques. The objective approach involves automatically collecting data about the application and its users (for example counts, sequence, timing of actions) including audio and video recording; automatic software monitoring; and psychological event monitoring (Hilbert & Redmiles, 1998). Problems with the objective approach include the resource intensive nature of interpreting feedback to extract key issues and that the context, which is vital in interpreting the meaning of the users' actions, is missing from the data (Hammomtree et al., 1994, Hartson et al., 1996). In contrast to this, a key benefit of the subjective approach is the ability to capture aspects of the users' needs, thought processes and subjective experiences. However, problems with the subjective approach do exist, particularly if the intention is to collect usability issues and not merely list software bugs (Hartson et al., 1996).

During 1999 a team working at the OU developed the Low Cost Remote evaluation method (LCR method) (Dunkley et al., 2000) for the evaluation of high fidelity prototypes linked to the subsequent redesign process. The LCR method is itself a remote adaptation of an evaluation method proposed by (Smith et al., 1999), that incorporated structured sessions between designers and users derived from contextual inquiry. In effect, rather than the verbalisation of their experience with a (semi-)working prototype, a user located remotely will record their experiences through a commentary, prompted by a questionnaire, and written concurrently. The original case study, described in (Dunkley et al., 2000), held that LCR suffered from a

> *"temptation for users to explore and interact before they have completed a written answer"*

with a concomitant loss of the conversational nature of the face to face designer-user sessions. This paper investigates whether re-establishing a conversation between user and evaluator leads to a more effective LCR method. The cost of this loss can then be assessed.

The paper is structured as follows. Section 2 describes the prototype evaluated in the study. Section 3 summarizes the LCR method, and discusses a number of issues raised in a previous study (Dunkley et al., 2000) that have motivated the investigation reported in this paper. Section 4 describes the new remote evaluation methods based on the deployment of Internet conferencing tools. Section 5 reports on the findings and Section 6 discusses issues of methodology as well as practicalities. Finally, Section 7 concludes the paper.

2 The case study

The case study we describe involves the Open University (OU). The OU is the major provider of distance-learning education at university level in the UK and Europe. Increasingly, the OU has moved to the electronic provision of its teaching services to students via the Internet and e-mail systems. A component of the course materials, the continuous assessment system, provides the major vehicle for driving the student's distance learning experience: students complete assessments as they study a course. The scripts are marked by associate lecturers (aka tutors) located remotely both from the student and the university campus. Marked scripts are returned to students for feedback on their progression, via the OU.

The move to electronic provision has been facilitated by the development of software for script marking, namely the eTMA (electronic Tutor Marked Assessment) Marking Tool. The original design of the eTMA Marking Tool was subject to conflicting requirements. In operation, the software should require the minimum of support and technical backup. The usage pattern would be one of fairly long gaps followed by intensive use for short periods. Visibility and affordance of design were therefore crucial issues.

The eTMA Marking Tool prototype consisted of four windows, which worked in conjunction with an MS Word document displaying the student's work. The most complex tasks were associated with the 'Score Allocation' window of the tool, where the user is required to enter the score followed by feedback for each question. In contrast, when marking on paper, this order is not enforced. This design rationale came about because a score is a necessary component of the marking, whereas the comment is optional: the user might forget to add the score if the order were reversed. However, the importance of this action sequence needed to be conveyed to the user. An important issue was the visibility of the comments. With a previous version of this software it had been discovered that extensive comments could disrupt the format and display of the e-TMA returned to the student. In this prototype the comments were embedded so that they could be seen when returned to the student but disappeared as far as the user was concerned, although they could be glimpsed as the cursor passed over the score in the Word document.

3 The LCR Method

The LCR method attempts to capture the user's response to prototype interfaces in a contextual manner and provides a framework to simulate a remote conversation between the developer and the user. The method was strongly influenced by the ideas of contextual inquiry (Holtzblatt & Jones, 1993). There are a number of evaluation methods that are variously known by the terms think-aloud, verbal protocol and cooperative evaluation. Many experts recommend thinking-aloud for most ordinary face-to-face applications, although (Goguen, 1996) criticises such methods as 'unnatural'. Co-operative evaluation is a variation of think-aloud in which the user is encouraged to see himself as a collaborator in the evaluation rather than just a subject. This is claimed to be less constrained and the user is encouraged to actively criticise the system by the evaluator who is not necessarily the designer (Wright & Monk, 1991). We were interested in developing remote evaluation methods that could simulate this situation. The approach is based on integrating contextual enquiry approaches with simulated think-aloud methods with the

particular aim of promoting developer-user conversations, which the developer does not dominate. A key part of the method is the establishment of a conversation between the user and the developer supported by a series of questions structured within Norman's seven stages of action. The process is as exemplified in Figure 1 and 2, with reference to the eTMA prototype evaluation.

For each task
Ask the user to explain what s/he is attempting
For each sub task
Ask the user to explain what s/he is attempting
For each stage in Norman's model of interaction
Consider asking a question from the checklist
Next stage
Next sub task
Next task

Figure 1 Eliciting user comments in an LCR Session.

Norman's Stages	*Remote Evaluation Questions*
Form a goal	How does the screen help you select a way of achieving your task?
Form an intention	What is the most important information visible when you start to allocate the score and make comments?
Specify the action sequence	How does the Score Allocation window make it obvious how to allocate scores and make comments?
Execute action	
Perceive the resultant system state	How has the Score Allocation window changed in order to show what you have achieved?
Interpret the resultant state	How do you know what you have done is correct
Evaluate the outcome	How would you recognise any mistakes? What action would you take to correct any mistakes?

Figure 2 LCR evaluation framework: sample questions.

The LCR method applies these concepts to enable users to articulate the way in which they would use a prototype interface to complete their normal tasks. Unless the evaluation focuses on specific tasks and context, users tend to evaluate prototypes in abstract terms referring to their general view of the interface and about whether they like the font, colours, etc. Users may not recall problems with the interface outside the context of actually doing work.

In general, remote evaluation experiments reported in the literature have taken place in organisational settings. There, a single group of remote users were observed in their normal work environment, which could, to some extent, be controlled and where audio and video equipment could be set up (Holtzblatt & Jones, 1993). In contrast the e-TMA Marking Tool experiment reported in (Dunkley et al., 2000) involved users at widely distributed locations. Their social context was the home, with background family noise. Video recording and video conferencing were not feasible options, as most of the users would not have the equipment. Asynchronous evaluation facilitated the experiment due to the time distribution of the users' work patterns. Consequently, LCR was designed to simulate the 'think-aloud' method by providing an electronic, user-completed journal to capture users' responses during interaction with the prototype. Hence the name 'write-along.

Additionally, the LCR method consisted of an evaluation package with the following components:

1. detailed evaluation form;
2. critical incident report form;
3. summary form of nine open-ended evaluation questions.

To summarise the conclusions of the LCR study in (Dunkley et al., 2000):

- The remote evaluation was effective in providing information from which the developer team could identify usability problems that lead to design changes.
- Users were able to articulate their experience of using the interface, regardless of their gender and task complexity.
- Users needed to get used to the conversational style of the LCR framework; repetition of the style of questions for each task assisted this.
- Since there is some evidence of a learning effect in terms of the questions being asked the evaluation design should ensure less complex tasks are encountered at first.

It was also found that although users were able to understand the concepts of critical incidents (Castillo et al., 1998) and report these effectively, few usability problems were identified in this way.

The LCR method itself was highly efficient in terms of the resources needed to extract the usability problems from the users' responses. The one element that was a problem in the write-along LCR method which cannot arise in think-aloud protocols is that users get tempted to explore and interact before they have completed the written answer. An actual observer can prevent this in a way that is difficult to simulate in the LRC remote evaluation. We therefore considered that the LCR method needed further development and tool support by the incorporation of audio prompts or active agent technology to maintain the conversational nature of the evaluation. This paper describes the further work that has been carried out. In particular, it describes an investigation into ways to combine the write-along method with audio prompts and reminders. This was carried out by adapting the LCR method for use with two conferencing systems, Lyceum and NetMeeting™, as described in the following section.

4 Real-time conferencing investigation

We have adapted the original LCR package based on a write-along method to include audio prompts designed to keep users to the task scenario and to help any

who experience critical incidents. Additionally, the evaluator and user can discuss a critical incident immediately instead of the user reporting them for later analysis (Hartson & al., 1996).

We describe an experiment to assess any increase in effectiveness of extending the LCR with real-time conferencing. The conferencing packages used to extend the LCR were Lyceum and NetMeeting™. The additional functionality provided by Lyceum was many-to-many voice and data conferencing. That of NetMeeting™ was to allow the evaluator to view the user's desktop at the same time as peer-to-peer voice conferencing.

We used the same prototype of the eTMA Marking Tool as in (Dunkley et al., 2000), as its usability problems had already been identified by that study, and through subsequent conventional usability evaluations as well as actual implementation.

The set of users was drawn from the same population of associate lecturers as those of (Dunkley et al., 2000). This is a large population, so we were able to select users with no previous experience of the electronic assessment tool and check that their user profile was comparable to that of (Dunkley et al., 2000). By and large, the users were new to Lyceum and NetMeeting™. We used the same question and answer framework (see Figure 2), but the questions were adapted to a briefer style needed for audio communication over the Internet.

We also used the same task scenario, which consisted of four tasks:

1. selecting courses and scripts to mark;
2. setting part marks for the standard mark scheme;
3. marking sample scripts;
4. storing completed marked scripts.

Users were asked to complete tasks 1. to 4. online, then explore further completion of the tasks off-line, using the LRC write-along method, and complete a questionnaire.

4.1 Lyceum Trials

Lyceum is a voice groupware system developed at the OU, designed to support real-time collaborative eLearning (Rapanotti & Hall, 2000). Lyceum users can participate in a real-time voice-based conference and share collaborative tools.

Lyceum is a client/server system. Users install the Lyceum client on their PC and access the Lyceum server across the Internet via an ISP or a corporate LAN. The Lyceum client is designed to run on mid-range Windows PCs with standard multimedia support.

Among Lyceum's features, the following were particularly relevant to our investigation:

- Participants met in *virtual rooms*, hooked by the Lyceum server. In each room they took part in real-time collaborations and discussions. In particular, they could talk to each other and share *collaborative tools*.
- Two collaborative tools deployed with Lyceum were used during the sessions: a whiteboard and a text editor. The whiteboard is a simple generic drawing application that allows participants to sketch simple shapes and annotate them with text. The text editor allows users to edit a text file collaboratively.

- The Lyceum *voice tools* (see Figure 3) facilitate the moderation of plenary audio sessions. These include: a 'Talk' button, that has to be kept pressed when talking; a simple voting system ('Yes/No/Wipe' votes); and a request-to-speak tool (hand up).

Figure 3 Lyceum's voice tools.

- A *text chat channel*, for synchronous text communication, was used as a back-up channel in case of audio failure.

4.1.1 The sessions

The Lyceum sessions were designed to allow a number of evaluators and users to work together to perform remote usability evaluations synchronously. We divided our users into groups, each group made up of 4 to 5 people. Each user group took part in one training and one evaluation session (see below) for a total of just over two hours online. Training and evaluation were structured as follows. Each started with a plenary session for all participants, followed by individual activities in breakout rooms, followed by a plenary discussion to close. Each session was run by two evaluators.

In addition to evaluating the extended LCR method, we wanted to exploit the many-to-many voice and data conferencing facility of Lyceum by ending the individual remote evaluations with a plenary discussion of the prototype's usability. During these plenary activities, one of the evaluators would moderate and record the discussion, while the other would deal with technical problems or latecomers. During individual evaluation sessions, the two moderators would visit the users' breakout rooms in a round-robin fashion. The users were also completing write-along documents as they worked through their tasks. During the plenary session two collaborative tools were used: a document editor and a whiteboard. The document editor was used to exchange evaluation information between evaluators and users; the whiteboard to record comments during a plenary discussion.

Prior to the evaluation sessions, we ran training sessions with the users to accustom them to Lyceum, voice moderation and voice communication protocols, as well as introducing them to the mechanics of the subsequent online evaluation sessions.

During all the sessions, the users were connected to the Lyceum server via an ISP, while the evaluators worked off a LAN.

4.2 NetMeeting™ Trial

In the development of the original LCR method the use of verbal protocols through telephone links was considered, but discounted as impractical due the disruption it would have caused to the users' work, and that it made communication with the evaluator, who could not see the interface, difficult.

In the experiment, we adopted a face to face method described in (Wright & Monk, 1991) for use with NetMeeting™, which we use to link the evaluator and user. NetMeeting™ is an audio, video and data conferencing system developed and distributed by Microsoft© Inc. NetMeeting™ supports real-time conferencing and allows sharing users' desktops across the Internet[1].

4.2.1 The sessions

In each NetMeeting™ session, the two participants, one evaluator and one user, were connected peer-to-peer in voice and data conference. This allowed evaluator and user to converse and share collaborative tools synchronously. The user's desktop was shared during the sessions allowing the evaluator to perceive events occurring on the user's PC. The evaluator, taking the role of the designer in the face to face method guided the user through the usability evaluation, by conducting a remote conversation. We made use of the NetMeeting™ file transfer facility in order to exchange tasks and evaluation documents between evaluator and user. The feedback loop was closed with the user's desktop being visible to the evaluator.

Difficulties with users' machine base and ISPs meant that it was not possible to obtain audio of an acceptable quality. (We note that similar problems were reported in previous experiments with low cost conferencing software, including NetMeeting™ (Shah et al, 1998).) Therefore, the NetMeeting™ sessions were conducted on a campus LAN with evaluator and user located in separate rooms. We made use of two high specification laptops, appropriately configured and tested for NetMeeting™ use. This allowed the same laptops to be used in all sessions. However, even in this configuration, the audio quality suffered excessive distortion.

5 Findings

Table 1 presents data extracted from the analysis of user comments. Column 1 contains the number of separate design issues/usability problems identified in the original LCR evaluation (Dunkley et al., 2000), together with the number identified in, respectively, the Lyceum and NetMeeting™ trials. Note that results for Lyceum include both those for users working online and subsequently working off-line independently.

R in the table refers to the average rating awarded to the usability problems by an independent HCI specialist. (Rating of the usability problems on a scale: 0 generally usable; 1 minor usability problem; 2 significant usability problem; 3 serious usability problem; and 4 catastrophic usability problem.)

Marking Tool Windows	**LCR** (13 users)	**Lyceum** (9 users)			**NetMeeting** (4 users)
		On-line	*Off-line*	*Total*	*Total*
Marking	10 (R= 3.5)	2 (R= 3.5)	3 (R= 3.6)	5	2

[1] Unlike Lyceum, NetMeeting™ does not support virtual rooms and many-to-many voice (at least as used in peer-to-peer mode).

Scheme					
Main Window	7 (R= 2.9)	2 (R= 3.0)	3 (R= 3.0)	5	3
Score Allocation	17 (R=3.4)	9* (R= 3.8)	4 (R= 3.0)	13	5
Total	34 (R= 3.2)	13 (R= 3.5)	10 (R= 3.5)	23	8

Table 1 Summary of Usability Problems Identified - * Includes four new usability issues, identified through the plenary session discussions, that had not been identified in the previous LCR method.

The experimental method adopted was based on a between-groups randomised design where different users were involved in the different sessions rather than the same users being used for all three sessions. The advantage of a between-groups design is that any learning effect resulting from the user repeating the test with the same interface is controlled. The danger of there being significant variation between the groups was handled by carefully selecting the users who were drawn from the same population of associate lecturers and completed a user profile questionnaire prior to selection. The same issues were identifiable from many comments in all the three studies. Table 2 summarises a detailed analysis of user groups versus method of evaluation. For example, with the LCR method as described in (Dunkley et al., 2000), the 13 remote users identified 34 separate usability problems. However, from the table, there is no significant difference between the three methods in terms of the mean or standard deviations of the number of usability problems identified by the different evaluation methods. In addition the difference in means was tested in each case by using a *t* test. In both cases the differences were not significant and the null hypothesis accepted: Lyceum compared with the LCR gave a value for *t* of 0.38, and for NetMeeting™ compared with the LCR results, a *t* value of 0.53 was obtained. Both results are well below the critical values. Further analysis using ANOVA (one way) gives $F_{(2,23)}$ as 0.60, which is not significant, supporting the hypothesis that in terms of number of usability problems identified there is no significant difference between the three methods.

	LCR	**Lyceum**	**NetMeeting**
Number of users	13	9	4
Total of usability problems	34	23	8
Mean of usability problems	6.23	7.55	5.75
Standard deviation	3.76	2.78	1.89

Table 2 Statistical summary.

Although for the LCR- and Lyceum-based methods, there was no significant difference in the number of usability problems identified there was some difference in the nature of the usability problems. The Lyceum plenary sessions identified four usability problems not identified by the previous evaluation: although three had subsequently been identified by users after implementation, one was a new usability issue not previously identified. As can be seen from Table 1, the severity ratings in

the Lyceum trials are the same or slightly higher than those in the LCR trials, which implies that the more severe problems were identified in the Lyceum sessions.

For the LCR- and NetMeeting™-based methods, sessions with the 4 users located on campus seemed to identify fewer usability problems. Note that the sessions appeared more stressful due to the tendency for the audio to break up. Fewer tasks were completed in the online sessions and, post-evaluation, the users did not express the same level of satisfaction with these sessions. It was noted, however, that the ability to see the user's desktop was an advantage when users got into serious difficulties.

One interesting finding to emerge from the LCR study using the write-along method was that the majority of users' comments read convincingly as though written while looking and exploring the screen. Although the LCR method uses a question and answer framework (see Figure 2), it does so in a style to empower the user by simulating direct conversation with the remote developer on equal terms. (Holtzblatt & Jones, 1993) describes active engagement as having the sense of a stream of consciousness discussion and this feeling was recognisable in many of the users' responses (see Table 3). This indicated that the users' behaviour was not substantially interfered with by the constraint of writing along in a Word document as they carried out their tasks. In comparison, the comments were terser in the Lyceum sessions. This terseness can be explained as a response to increased time pressure. Table 3 gives examples of users' responses to the questions in the two methods - the write-along LCR method (users 1-3) and the Lyceum-based method (users A-C).

Question: *How does the Score Allocation window make it obvious how to allocate scores and make comments?*	
User 1: I am not sure it does. Zero in box*[sic]* made me experiment with putting score in and I discovered that if you clicked on arrow that this helped you position score in script.	User A: Each question or part of question is shown down the left-hand side of the score allocation. I hope this is the score allocation as it has only the name of the student for a title. The numbers are slightly confusing, as the question number and the question part are the same size.
User 2: There is a question tab at the top of the list of numbers for that question. The window shows a list of marks against a list of text fields. It seemed obvious to insert marks in the text fields and against each of the marks listed. To add a comment the arrow is raised and the text box is displayed whenever you scroll over it.	User B: The question numbers are listed.
User 3: Well, no I'm not sure. Your marking tool is confused. Blast. I pressed Yes for 'I am sure' and now it has put the mark in a silly place in the script.	User C: Not really at first. Problem: cursor was at 'start of text' but I could not move it because I could not move the score allocation window: each time I tried the error/warning window about the cursor

	position came up and became the active window preventing me from dragging the score allocation window away from the ...

Table 3 Samples of users' write-along responses during LCR (to the left) and Lyceum (to the right) sessions.

Lyceum users' perceptions were also captured in the post-evaluation questionnaire. It is evident that the users found the experience interesting and worthwhile. They soon got the hang of using the voice communication tools to participate effectively in the plenary discussion and spoke naturally in terms of their own virtual evaluation room and the plenary room. Sample users' views are set out in Table 4.

Question: *Were you comfortable communicating with the other users and the evaluators through Lyceum?*	
User 1.2	Very - once got to recognise voices - at first hard, as had to keep looking at who was talking. Easier 2nd time, as more familiar
User 1.3	Yes, except that I found at times that the sound reverberated in the earpiece making it difficult to distinguish what was required. I was comfortable with typing responses.
User 1.5	Very - it appears to be a splendid communications device.
User 2.2	Yes; this seemed more 'human' than simple text conferences; 'hand' seemed a little 'school-like' but its purpose was clear; yes/no flags seemed helpfully basic.
User 2.3	Apart from the connection problems this was OK
Question: *Did you feel any anxiety about the way the evaluation was set up?*	
User 1.1	Not anxiety as such: rather increasing consciousness that I was too unfamiliar with the context of the evaluation to make a very full contribution in the online trial: this was my first experience both with Lyceum and of electronic marking and a marking tool.
User 2.1	Yes. I would have done it differently. We were asked to install the marking tool, but not to run it. It is simple to navigate and I would have felt more at ease in the initial stages of the test if I had been allowed to 'play' with the tool and to examine its components etc. There are no great technical demands with it.
User 2.2	Not sure I'd use the term 'anxiety', I will admit to a little bemusement, and the feeling things other than an electronic marking tool were being assessed

Table 4 Extracts from users' reflective questionnaires.

The main advantage of using NetMeeting™ was the facility to share the user's desktop with the evaluator. The prototype interface was based on four interacting windows that were related to the different tasks. At least one of these users was observed trying to complete a task using the wrong window. This had also happened when the LCR write-along method had been used in (Dunkley et al., 2000), but, because the desktop was not visible to an evaluator, the user was not able to correct

the error for some time. The poor and unpredictable audio quality made the sessions very taxing for the user and evaluator. It was noticeable from the transcripts that both participants needed to frequently seek confirmation that what they had said had been received and understood.

6 Reflection and Discussion

In our study, we have adapted the LCR method - a text based asynchronous remote evaluation method - for use with real-time audio and other synchronous conferencing tools over the Internet. The aim of the study was to evaluate the feasibility and effectiveness of this adaptation, compared with the original LCR. In the previous section, we have reported some of our findings related to the usability defects captured during the trials, as well as some users' perceptions of the trials. In this section, we reflect on methodological issues as well as discussing some of the practicalities.

6.1 On the methodology

Is it sensible to ask whether our comparison is valid or, because of the experimental conditions and variations between the experiments, we were actually comparing 'apples and pears'. Most usability evaluations are focused on evaluating differences in the user interfaces. In this case we use the same interface and the experiment was designed to investigate the effectiveness of the different variations in the evaluation methods.

6.1.1 User profile

Each experiment employed different users, but they were samples form the same population of associate lecturers. Every user completed a user profile questionnaire prior to selection.

6.1.2 Prototype

The same prototype interface was the subject of each experiment and so each experiment had the same number of extant errors.

6.1.3 Method/treatment

Between the NetMeeting™ and LCR trials, there were differences in the methods of evaluation. However, these differences were limited to using written questions versus having an evaluator remotely monitoring and questioning. The main methodological ramification of this difference was that, since the evaluator could see the user's desktop, s/he could help during a critical incident rather than through an asynchronous communication (email exchange) which interrupted the evaluation. The methods of counting/measuring discovered usability problems were the same.

The methods of evaluation of Lyceum and LCR were more dramatically different because the Lyceum experiment involved users interacting as a group. Again, the methods of counting/measuring the usability problems were the same

The variances between the different experiments is detailed in Table 5.

	LCR	**Lyceum**	**NetMeeting**
Tool evaluated	Marking tool prototype	Marking tool prototype	Marking tool prototype
User training with tool to be	none	none	none

evaluated			
User training with conferencing tool	Not required	Testing to check equipment and connection; training session for audio moderation and collaboration	Training session to check equipment and connection
Number of users	13	9	4
Total usability problems identified	34	23	8
Task description and evaluation questions	in separate Word document emailed prior to evaluation	in shared Lyceum text editor available during the session	in separate Word document, transferred during the session
Documents sharing	none	yes, through Lyceum text editor	none, but user's Word document visible through desktop sharing
Type of session	Asynchronous, offline. Individual user	Synchronous, online. With 4-5 users and two evaluators. Combination of individual and plenary activities	Synchronous, online. Individual user with peer-to-peer communication with evaluator
Contact with evaluator	No direct contact	2 evaluators per group of 4 - 5 users	One-to-one evaluator contact
Audio prompts	none	yes for plenary work; LCR for individual work with occasional audio prompts	yes for task 1, followed by LCR
Group work	none	Plenary discussion with group	none
Evaluator's visibility	Evaluator only sees screen shots after evaluation completed	Evaluator cannot see desktop, but can see shared task document	Evaluator sees users desktop at all time

Table 5 Summary of the experiments.

6.2 On the practicalities

Several factors may affect the effectiveness of a remote evaluation method based on Internet conferencing tools. Among those, the following issues were exposed during our investigation. Some deserve further investigation and future experiments will be designed to address some of those issues.

6.2.1 Session duration

Online evaluation sessions, whether with Lyceum or NetMeeting™, can make high demands on the participants in terms of concentration, given that they are required to carry out activities in real-time, with visual and audio input and output occurring at the same time. This effect is compounded by, for example, the often imperfect audio quality, instability of Internet connections and variations in traffic thereon. We found that a continuous online session of one hour was optimal.

6.2.2 Session scripts

For each session, evaluators were given designed session scripts. Each script provided a detailed plan of the session, including the set of activities, their sequencing, the tools required, and prompts to help the evaluators moderate the session smoothly. The scripts were very valuable for the evaluators during the sessions, acting as memory aids and helping with time keeping. They also allowed the rehearsal of the sessions and the running of comparable sessions with different groups of users.

6.2.3 Beta-testing

We beta-tested all our sessions in-house, before going live. This allowed us to test and finely tune all session designs. Moderating a synchronous online session is a sort of live performance and some general rehearsal is necessary, in particular in the Lyceum's many-to-many scenario. For instance, beta-testing our initial design for the Lyceum evaluation session highlighted the need for familiarising the users with Lyceum prior to the evaluation session, hence the introduction of training sessions. Also, it showed that it was very important to put users into separate virtual rooms. Although each user could work on their own document, even in a plenary room, it soon became evident that users were conscious of their progress versus that of other users and this caused them anxiety if they felt they were experiencing difficulties other users did not, or were making slower progress.

6.2.4 Online and offline mix

Compared to the LCR method, the online sessions prevented the users from becoming lost during the evaluation tasks - for instance, trying to perform a task in the wrong window. This points to strength of the online methods over the LCR. On the other hand, working online can be quite slow and, as mentioned above, can be tiring. This limits the amount of work that can be carried out during a single session and makes online working unsuitable for usability evaluation of complex software interfaces. Our trials seem to indicate that there is a trade-off to be made: a combination of online and off-line work should be adopted; neither solely online or solely off-line sessions are optimal.

Also, in particular during NetMeeting™ sessions with desktop sharing switched on, some of our users felt under constant observation and evaluation, which added to

pressure and anxiety. In contrast, as reported in some of the users' answers in Table 2, Lyceum users liked the chance of discussing their ideas online in plenary discussions, and regarded online collaboration a positive part of the process.

6.2.5 Technical support

The amount of technical support required deserves serious consideration. In the current Internet panorama, it is unlikely that any real-time conferencing software could be deployed without technical support, in particular with real-time voice and collaboration. Some technical support will be necessary at least for software installation and proper configuration of the end-users' PCs. More realistically, further support may be required for fine-tuning of the software, to cope with heterogeneous PC settings, and the volatility of today's Internet connections.

6.2.6 Recording facilities

We captured the audio of all sessions for research purposes and for the record. Audio capture was accomplished simply by using a high quality tape recorder and microphone connected to one evaluator's speakers. Digital audio and video-capturing software could not be run on any of the participants' PCs because both Lyceum and NetMeeting™ require the exclusive use of a PC's sound card. Also, such programs tend to interfere with the conferencing software operation and cause a degradation of their performance. We are aware that this is not a perfect solution, and better methods may be required for larger scale trials.

7 Conclusions

The paper has reported on an investigation into the development of an effective remote evaluation method applicable to users in their natural working environment. The method leverages Internet communication and collaboration technology to facilitate the conversation between user and evaluator.

By comparing the two online approaches taken in the investigation, we found that the original LCR method was usefully extended by the addition of voice conferencing. The Lyceum tool provided this, and appears to have the potential for further development. In particular, the software provided a range of facilities that could support task based and focus group based evaluation. In contrast, the ability to share the user's desktop was strength of NetMeeting™, and proves a great advantage when dealing with critical incidents.

Acknowledgement

We thank our colleagues at the Open University, in particular the Lyceum development team for all their support, and the LTIC for funding this project.

References

Carroll, J.M. & Rosson, M. B (1991), “Stalking the View Matcher”, *Human-Computer Interaction*, **6**, 3 and 4, 281-318.

Castillo, J.C., Hartson, H.R., & Hix, D. (1998), “Remote Usability Evaluation: Can users report their own critical incidents?”, *Proceedings of CHI'98*, ACM Press, 253-254.

Dunckley, L., Taylor, D., Story, M., & Smith, A. (2000), “Low Cost Remote Evaluation for Interface Prototyping”, *People & Computers*, XIV, 389-404.

Goguen, J. A. (1996), “Formality and informality in requirements engineering”, *Proceedings of ICRE'96*, IEEE Computer Society Press, 102-108.

Hammontree, M., Weiler, P., & Nayak, N. (1994),“Remote Usability Testing”, *Interactions*, July, 1994, 21-25.

Hartson, H.R., Castillo, J.C., Kelso, J., & Neale, W.C. (1996), “Remote evaluation: the network as an extension of the usability laboratory”, *CHI'96, Proceedings of ACM*, 228-235.

Hilbert, D.M., & Redmiles, D.F. (1998), “An approach to the Large-Scale Collection of Application Usage data over the Internet. Separating the Wheat from the Chaff”, Internet-mediated user feedback, *Proceedings of the Workshop on Internet-based Groupware for user participation in product development*, 1998.

Holtzblatt, K., & Jones, S. (1993), “Conducting and analysing a contextual interview”, Schuler, D., & Namioka, A. (eds.), *Participatory design: principles and practices*, Lawrence Erlbaum Associates, 177-210.

Rapanotti, L., & Hall, J.G. (2000), “Lyceum: the system and its architecture”, *Proceedings of ED-ICT2000,* Vienna, 43-52.

Shah, D., Candy, L., & Edwards, E. (1998), “An investigation into supporting collaboration over the Internet”, *Computer Communication*, **20**, 16, 1458-1466.

Smith, A., Dunckley, L. & Smith, L. (1999), “Importance of Collaborative design in Computer Interface design”, M.A.Hanson, E.J.Lovesey & S.A.Robertson (eds.), *Proceedings of Contemporary Ergonomics ’99*, Taylor & Francis, 494-498.

Wright, P.C., & Monk, A. F. (1991), “A cost-effective evaluation method for use by designers”, *International Journal of Man-Machine Studies*, **35**, 6, 891-912.