Cueing Retrospective Verbal Reports in Usability Testing Through Eye-Movement Replay

Nicola Eger
Sony Ericsson Software Applications Laboratory
Warrington, WA3 6AE, UK
+44 (0)1925 281556
nik_eger@hotmail.com

Linden J. Ball
Department of Psychology
Lancaster University
Lancaster, LA1 4YF, UK
+44 (0)1524 593470
L.Ball@lancaster.ac.uk

Robert Stevens, Jon Dodd
Bunnyfoot
Harwell Innovation Centre
Harwell, OX11 0QG, UK
+44 (0)845 6440650
{rob, jon}@bunnyfoot.com

ABSTRACT
An experimental validation is presented of a novel method for usability testing that entails the playback of dynamic eye-tracking data to cue the elicitation of retrospective verbal reports. Participants in our study produced: (1) think-aloud reports during an online search task, and (2) retrospective reports during another online search task, with reports being cued by the playback of either the screen capture of events or the participant's own eye-movements. Task-completion times and response rates were recorded for all reporting methods. Fewer participants completed the search task whilst thinking aloud, indicating the reactivity of this technique. Verbal transcripts were coded for instances of usability problems. The eye-cued method identified more usability problems than the think-aloud or screen-cued methods. A significant interaction between search engine type and retrospective cue type suggests that the value of the eye-cue method for eliciting usability problems may be greatest with more complex search environments. Our results demonstrate that when cued appropriately, retrospective reports may be less reactive and more informative than other verbalisation techniques.

Categories and Subject Descriptors
H5.2 **[Information Interfaces and Presentation (e.g., HCI)]**; User interfaces - *evaluation/methodology.*

General Terms
Human Factors.

Keywords
Eye-tracking, verbal protocols, online usability testing.

1. INTRODUCTION

1.1 Think-Aloud Reports
The concurrent 'think-aloud' method is deployed widely in usability research and requires people to verbalise their thoughts whilst performing tasks [2, 4, 14, 23, 24, 28]. The technique was developed by Ericsson and Simon [15] in line with their information processing model of human cognition, which proposes that cognitive processes entail a sequence of internal states that are transformed by mental operations. A key assumption of this model is that mental states and operators are active in short-term memory, such that they are directly available for verbal reporting. As such, if a participant (e.g., a system user) concurrently verbalises their thoughts during task performance then the researcher or usability analyst can gain insights into the nature of ongoing cognitive processes, including difficulties being encountered and strategies being invoked to deal with them. Although this technique has high face validity it comes with certain limitations, as discussed in the following sections.

1.1.1 Inaccuracy
Nisbett and Wilson [22] questioned the accuracy of verbal reports. They argued that observing one's own mental processes does not provide any special insight into cognition because people generate 'causal theories' about their own behaviour, rather than consulting a memory of the mediating processes that actually determined their behaviour. Nisbett and Wilson suggest that verbal reports will be most accurate when influential stimuli are still present in memory that are also plausible, and when few plausible but non-influential factors are available. Ericsson and Simon countered that think-aloud reports or 'protocols' are not liable to the inaccuracies suggested by Nisbett and Wilson, since only the contents of short-term memory are being reported by participants when thinking aloud [16], freeing participants from the effort and unreliability of accessing an embedded long-term memory trace. Research, however, has indicated mixed results on this accuracy issue, which remains far from settled. For example, concurrent verbalisations of design activity have been shown to map poorly onto the behaviour that the descriptions were intended to characterise [12], whilst little effect on veridicality was evident in children's concurrent reports of subtraction processes [27].

1.1.2 Incompleteness
Evidence suggests that think-aloud verbalisations may be incomplete as some processes are unconscious [3] or difficult to translate verbally [31]. Participants can covertly edit their speech, choosing what to keep undisclosed, perhaps in response to social pressures to please the experimenter (e.g., a system user may refrain from reporting usability difficulties if they believe the researcher has a vested interest in the system under test), or because they think that certain information is too obvious to be mentioned. Finally, people think quicker than they talk, such that only a sample of cognitive activity may be reported [3].

© Nicola Eger, Linden J. Ball, Robert Stevens, Jon Dodd, 2007
Published by the British Computer Society
People and Computers XXI – HCI… but not as we know it: Proceedings of HCI 2007
Linden J. Ball, M. Angela Sasse, Corina Sas, Thomas C. Ormerod, Alan Dix, Peter Bagnall, and Tom McEwan (Editors)

1.1.3 Interference

Thinking aloud can also interfere with a primary task by changing the normal strategies that are deployed. Such 'reactivity' can be *negative*, disrupting natural thought processes in a detrimental manner (e.g., slowing solution times and inhibiting task completion), or *positive*, for example, facilitating task completion. Considerable evidence demonstrates negative reactivity across many domains [e.g., 12, 13, 29, 30]. Particularly striking is the 'verbal-overshadowing' effect [30], which is revealed as markedly worse performance under verbalisation requirements for tasks involving essentially 'non-reportable' cognitive processes. The beneficial effect of thinking aloud on certain tasks is also well documented [e.g., 1, 6, 10, 17], and may arise from metacognitive processing linked to the production of explanations [5]. Altered task performance clearly runs counter to Ericsson and Simon's claim [16] that thinking aloud should not affect primary-task processes. How the experimenter interacts with the participant can also interfere with normal task processes; even neutral comments can taint subsequent performance by re-directing attention [20]. For this reason Ericsson and Simon [16] recommend the use of only a non-directive prompt to "*Keep talking*" if the participant becomes quiet - otherwise the researcher should not intervene. However, in the context of usability testing, practitioners typically do not give instructions in the prescribed manner, and, instead, use prompts such as "*What do you think this means*?" [7, 14]. Such prompts can disrupt the flow of task-based processing, redirect attention and invoke retrospective rationalisations.

1.2 Retrospective Reports

A less popular protocol method, retrospective verbalisation, can combat some of the limitations of thinking aloud. Retrospective protocols appear to have good validity if given immediately after the primary task, since relevant information is retained in memory and can be directly reported or retrieved via contextual cues [16]. The experimenter simply asks the individual to report the thoughts that they remember arising during the prior task. Interestingly, participants are found to be highly proficient at doing this [15, 34]. The main problem with the method, however, is its reliance on memory. Encoding, storage and retrieval processes are fallible owing to generalisations, interference and forgetting [16, 29, 33]. Retrospective reports are open to post-hoc rationalisations, bias and fabrication. Although concurrent verbalisations may suffer a similar fate [22] the risk is reduced as participants have little time or resources to fabricate or edit speech. Producing retrospective protocols can also be taxing as having to perform the task and then give a verbal account increases fatigue. Despite these drawbacks, retrospective protocols have advantages over concurrent protocols. Attentional resources are not strained during primary task processing, so there are no worries about reactivity. Moreover, since the cognitive system is also not under strain during the retrospective phase, the participant is free to verbalise naturally, consequently decreasing the amount of unfinished, incomprehensible statements the experimenter has to transcribe and make sense of. In addition, if probes are used during retrospection, there is no chance of altering behaviour as this has already taken place.

1.3 Think-Aloud versus Retrospective Reports

The limited research comparing think-aloud with retrospective protocols reveals little evidence of a difference between the two [e.g., 27], and the usability literature describes them as 'equal alternatives' [23]. In one study [18] retrospective reports, aided by watching computer-log files of activity, produced similar results to concurrent protocols. In another study [8], retrospective reports elicited during usability testing of multiple window-handling produced fewer verbalisations than concurrent protocols, and reporting tended to focus on explanations rather than procedures. However, this study was limited as it did not test the type of problems reported, a key aspect of usability testing. Van den Haak et al. [34] aimed to assess the validity of concurrent and retrospective protocols during usability testing of an online library catalogue. The two techniques showed comparable results concerning the quantity and quality of usability problems identified, but the way in which the problems were highlighted differed: retrospective protocols revealed problems verbally whilst in the concurrent approach problems emerged via non-verbal behaviours (e.g., expressions) and on-screen actions. Thinking aloud also showed negative reactivity despite its rigorous application.

1.4 Eye Tracking and Thinking Aloud

Eye-tracking provides quantitative data about a participant's cognitive processes, based upon the 'eye-mind hypothesis'. This assumes a strong correlation between where an individual is looking and what they are thinking about [25]. Increased fixation duration is often taken to reflect increased cognitive demand or confusion, whilst processing difficulties may produce patterns of repetitive fixations or fixations located close together. Eye-tracking in usability research allows the focus of visual attention to be recorded so that any difficulties the participant has with a display can be inferred [e.g., 11, 19, 21, 26]. One problem with using eye-tracking as a stand-alone technique, however, is that it does not provide direct access to a participant's thoughts, feelings and experiences [25]. For example, in website usability it is difficult to distinguish whether an increase in fixation duration is due to cognitive demand or simply because the participant found some screen-based element particularly interesting [11]. Eye movements show the experimenter *where* the participant looked, but not *why*, revealing little about a user's intentions.

Think-aloud reports can qualify eye-tracking data by providing access to a user's thought processes. One particularly novel approach to combining eye-tracking data and verbal reporting in usability testing is to play back a person's own dynamic eye-movement trace to cue the elicitation of retrospective verbal protocols that may uncover usability problems. This method has been adopted with success in commercial usability testing (e.g., Bunnyfoot Universality, UK). However, we are not aware of any prior research in the usability domain that has subjected this approach to rigorous empirical evaluation.

1.5 Aims of the Present Study

Our study aimed to examine the validity of retrospective verbal reporting cued by eye-movement replay in a web-based usability context. It was hypothesised that the presence of the eye-tracking data would increase the accuracy and completeness of the retrospective report, thereby increasing the quantity (and possibly the quality) of usability problems highlighted. Such benefits were expected to arise from the combination of minimal cognitive load placed on the participant, the increased time for them to compose their thoughts, plus the affordance of direct access to where they had been looking on-screen during the given task. To validate the 'retro eye cue' method we set up an experiment that pitted its efficacy against two other verbal reporting procedures: (1) a standard think-aloud procedure whereby the participant simply

verbalised concurrent to primary task performance, and (2) a retrospective reporting method whereby the participant provided a verbal account cued by the playback of the dynamic screen events (including cursor movement) that had arisen during primary task performance (i.e., a 'retro screen cue' method). This study also aimed to assess the reactivity effects associated with thinking aloud; latency data alone are not a guaranteed index of reactivity, and so task completion rates were also measured [16].

2. METHOD

2.1 Participants

Twenty-four participants were tested (mean age 26 yrs; 50% male; all native English speakers). Participants were recruited via opportunity sampling from an internet café. All were regular users of the internet, and none had previously participated in user-testing research. They each received £10 for volunteering.

2.2 Apparatus

A Tobii 1750 remote eye-tracking system was employed that allows binocular tracking without impinging on the participant. ClearView eye-gaze software recorded eye movements and screen dynamics that could then be replayed in the retrospective conditions. Eye-movement data were superimposed onto visited webpages for the eye-cue condition (Figure 1). Camtasia was used to record participants' verbal reports.

2.3 Design

The experiment involved a nested design with mixed within- and between-participant factors (Table 1). The main factor was the Verbalisation Method: Think-aloud vs. retro eye cue vs. retro screen cue. Another factor was Search Engine: Participants completed a search task on two radically different search engines (i.e., Infomagnet and Google). Each participant produced a think-aloud protocol with one search engine and produced one of the two types of retrospective protocols with the other search engine.

Participants were divided into four test groups (Table 1). Task order was counterbalanced for each participant to prevent order effects. Dependent variables were: task completion time, task completion rate, quantity/quality of usability problems identified, and subjective measures from a questionnaire given to the participants.

Table 1. Experimental conditions.

	Think- Aloud Task		Retrospective Task	
Group	**Search Engine**	**Verbal Method**	**Search Engine**	**Verbal Method**
1	Google	Think Aloud	Info-magnet	Retro Eye Cue
2	Google	Think Aloud	Info-magnet	Retro Screen Cue
3	Info-magnet	Think Aloud	Google	Retro Eye Cue
4	Info-magnet	Think Aloud	Google	Retro Screen Cue

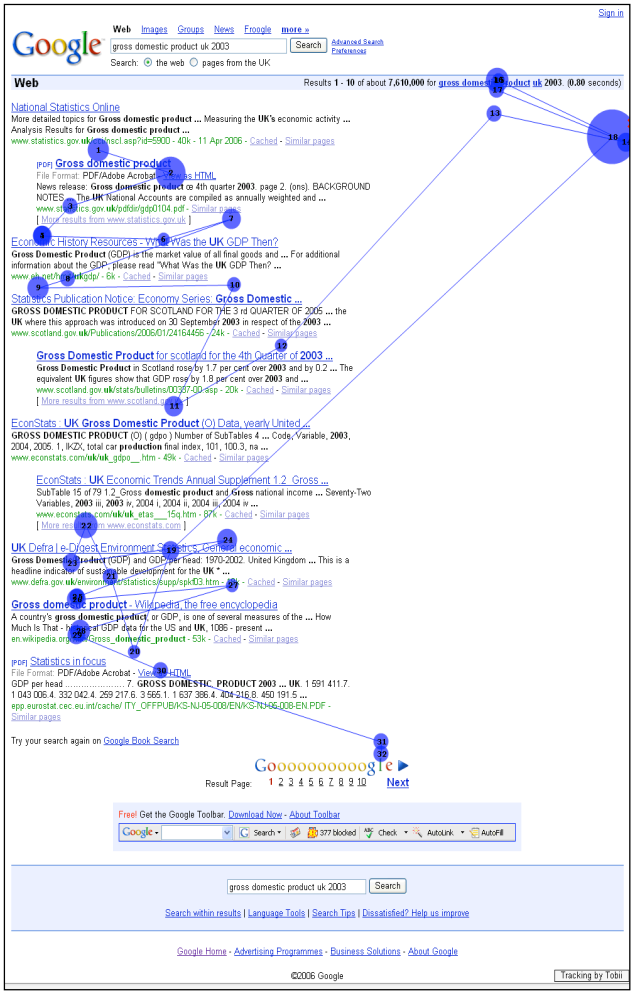

Figure 1. ClearView eye-gaze data replay. Blue dots represent fixation points that increase in diameter relative to fixation duration. The blue lines indicate the eye-movement scanpath.

2.4 Test Objects

Two online search engines were used: 'Infomagnet i-Globe' and 'Google Search'. Infomagnet is a novel tool employing the 'i-Globe', a moving visualisation of the earth that can be manipulated to display different aspects of data and to search for economic and geographic information. It consists of a user log-in area at the top of the page and a moving 'i-Globe' situated in the central space (Figure 2). The top panel is a search box, the middle panel allows manipulation of data, and the bottom panel controls how data are displayed on the globe. Once the options have been selected from the search box, the result is presented as a visualisation on the globe. It was thought that problems with the manipulation of the data panels, navigation of the globe and the aesthetics of the site would cause interesting usability issues.

Google is a familiar text-based search engine employing sophisticated algorithms to generate search results (e.g., based on site popularity). It consists of a search box situated at the top of the page with the results presented down the page displaying links to sites (see Figure 1). Sponsored links are presented to the right of the page, and the remaining results pages are accessible from the bottom of the screen. To control for the effect of variability in the search terms that could be entered and the consequent variety of results generated, only the second

page of the search engine was presented with the search terms predefined and a set page of results actually given.

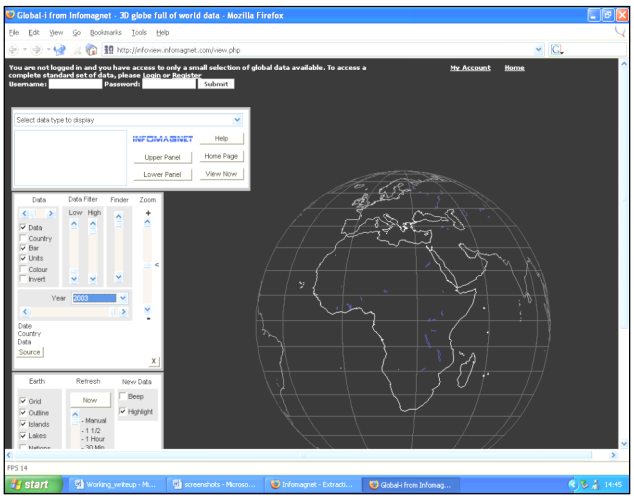

Figure 2. The Infomagnet i-Globe (accessible at infoview.infomagnet.com/view.php).

2.5 Questionnaires

A pre-experiment questionnaire elicited demographic information from participants (age, gender, occupation, education level) and ensured adequate internet experience. Post-task questionnaires based on previous research and involving five-point Likert scales [34] were employed to assess participants' views of the reporting methods. Questions were of three kinds: (1) overall experience of the method (to what extent was it difficult, unpleasant, tiring, unnatural, time-consuming); (2) how the method affected normal working (e.g., in terms of speed and focus); and (3) the effect of the experimenter's presence (unpleasant, unnatural, disturbing).

2.6 Tasks

The same search task was given to participants on each search engine: "*Find the GDP, annual growth percentage for the UK in 2003*". Participants were asked to alert the experimenter when they had found the answer. Each participant was given five minutes to complete the task.

2.7 Procedure

Participants provided one think-aloud protocol and one retrospective protocol (following task completion). If they fell silent for 15s the researcher intervened with questions from a pre-defined script [cf. 15]. Each participant was calibrated to the eye tracker and then given verbal instructions appropriate to the condition they were in:

- **Think Aloud:** "You will shortly be presented with a webpage. Using this site I would like you to find the GDP annual percentage growth of the UK in 2003. At the same time I would like you to tell me what you are thinking as you do this. We are testing the website not you, so any feedback we receive will be useful".
- **Pre-Retrospective for the Silent Task:** "You will shortly be presented with a webpage. Using this site I would like you to find the GDP annual percentage growth of the UK in 2003. I would like you to complete this task as you normally would, silently".
- **Retro Screen Cue:** "I'm now going to show you a playback of the task that you have just completed and I would like you to talk me through what you were doing and what you were thinking at the time".
- **Retro Eye Cue:** "I'm now going to show you a recording of your eye movements during the previous task and I would like you to talk me through what you were doing and what you were thinking at the time. The blue dots represent where you were looking; the longer you were looking at something the larger the blue dot becomes".

3. RESULTS

3.1 Usability Problems

Verbal transcripts were coded for instances of usability problems as categorised in previous research [28, 34]:

Layout: Visibility issues; failure to spot on-screen item; failure to absorb information; cluttering; irrelevant information/items; aesthetic problems.

Terminology: Failure to comprehend terminology of site.

Feedback: Application does not provide relevant feedback on actions or error messages; feedback not consistent with expectations; time issues of feedback.

Comprehension: User does not understand instructions, dialogue or actions of site.

Data Entry: User does not know how to conduct a search (enter search term, use drop down menu, start search etc.)

Navigation: Difficulty/failure to navigate around the page logically, or as desired.

Repeated-measures analysis of variance (ANOVA) tests were used to assess the quantity and quality of usability problems identified in the think-aloud and retro eye cue conditions (Table

Table 2. Mean usability problems identified (think aloud vs. retro eye cue) shown in bold typeface, with standard deviations in normal typeface, and *p* values for significant differences (ns = not significant).

	Usability Problems						
	Layout	Terminology	Data Entry	Comprehension	Feedback	Navigation	Overall
Think Aloud	**2.8** 2.8	**0.5** 0.6	**1.4** 2.3	**1.9** 2.7	**1.0** 0.9	**0.1** 0.1	**8.7** 4.9
Retro Eye Cue	**3.2** 2.0	**0.5** 0.8	**0.4** 0.6	**3.8** 4.4	**3.3** 2.1	**0.1** 0.2	**12.5** 6.8
Significance	ns	ns	ns	$p = .032$	$p = .012$	ns	$p = .014$

Table 3. Mean usability problems identified (think aloud vs. retro screen cue) shown in bold typeface, with standard deviations in normal typeface, and *p* values for significant differences (ns = not significant).

	Usability Problems						
	Layout	Terminology	Data Entry	Comprehension	Feedback	Navigation	Overall
Think Aloud	**1.8**	**0.5**	**0.5**	**2.5**	**2.4**	**0.2**	**8.8**
	1.3	0.8	0.7	2.7	2.15	0.36	4.77
Retro Screen Cue	**4.2**	**0.7**	**0.8**	**3.4**	**2.3**	**0.0**	**11.3**
	2.6	0.8	1.1	3.9	2.09	0.00	3.39
Significance	$p = .02$	ns	ns	ns	ns	ns	ns

2). The retro eye cue method generated more usability problems overall than the think-aloud method, $F(1, 11) = 8.52$, $MSE = 10.35$, $\eta_p^2 = .44$, $p = .014$. The nature of the problems identified also differed: Retro eye cue generated more problems relating to Feedback, $F(1, 11) = 8.91$, $MSE = 3.67$, $\eta_p^2 = .45$, $p = .012$) and to Comprehension, $F(1, 11) = 6.02$, $MSE = 3.35$, $\eta_p^2 = .36$, $p = .032$. The retro eye cue method seems, therefore, to have advantages over thinking aloud in identifying usability issues during online search.

A second set of repeated-measures ANOVAs examined differences in usability problems identified between the think-aloud and retro screen cue conditions (Table 3). Despite a trend for the retro screen cue method to elicit more problems overall (11.3 vs. 8.8), this effect was not reliable. There was, however, a significant difference in the quality of problems identified: Retro screen cue generated more Layout problems than the think aloud method, $F(1, 11) = 7.35$, $MSE = 4.77$, $\eta_p^2 = .40$, $p = .02$. It seems, then, that a retrospective protocol cued by the screen capture of events is generally neither more nor less productive in identifying usability problems than a think aloud protocol.

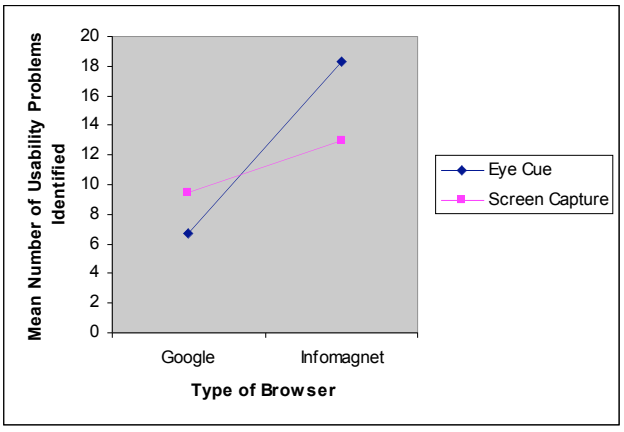

Figure 3. The influence of Search Engine (Infomagnet vs. Google) and Retrospective Cue (retro eye cue vs. retro screen cue) on the mean number of overall usability problems identified.

A series of 2-way between-participants ANOVAs assessed the effects of Retrospective Cue (retro eye cue vs. retro screen cue) and Search Engine (Infomagnet vs. Google) on quantity and quality of problems identified. The ANOVA on the overall number of problems identified revealed no effect of Retrospective Cue. There was, however, a reliable effect of Search Engine: Infomagnet generated more problems overall, $F(1, 20) = 35.54$, $MSE = 9.71$, $\eta_p^2 = .64$, $p < .001$. There was also a reliable interaction between Retrospective Cue and Search Engine, $F(1, 20) = 10.31$, $MSE = 9.71$, $\eta_p^2 = .34$, $p = .004$. The eye cue method was particularly good at detecting usability problems with Infomagnet, while the screen cue method was more beneficial for Google (Figure 3).

ANOVAs examining the specific *types* of problems identified also revealed no main effects of Retrospective Cue, but reliable effects of Search Engine on some problem types: Infomagnet was related to more problems of Data Entry, $F(1, 20) = 7.76$, $MSE = 2.60$, $\eta_p^2 = .28$, $p = .01$, and Comprehension, $F(1, 20) = 60.69$, $MSE = 4.62$, $\eta_p^2 = .75$, $p < .001$, whilst Google generated more Layout problems, $F(1, 20) = 5.29$, $MSE = 4.53$, $\eta_p^2 = .21$, $p = .032$. Furthermore, there was a significant interaction between Retrospective Cue and Search Engine for the average number of Feedback problems identified, $F(1, 20) = 4.93$, $MSE = 3.73$, $\eta_p^2 = .20$, $p = .038$, with the retro eye cue method again offering an advantage for the reporting of Feedback problems with Infomagnet, while the retro screen cue method had a particularly negative effect (see Figure 4). The opposite was true for Google: retro screen cue enhanced identification of Feedback problems.

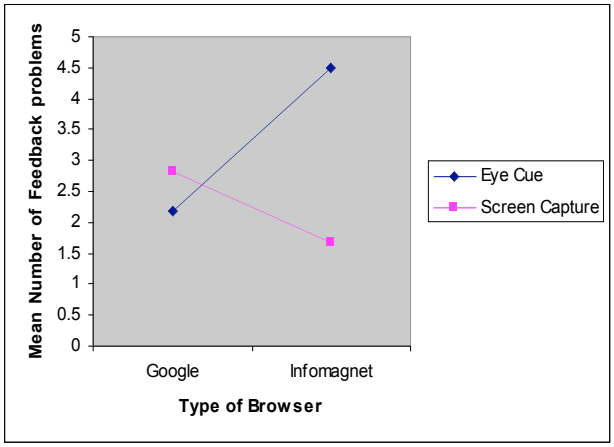

Figure 4. The influence of Search Engine (Infomagnet vs. Google) and Retrospective Cue (retro eye cue vs. retro screen cue) on the mean number of feedback problems identified.

A Search Engine effect was also evident in think-aloud conditions, confirmed by between-participants ANOVAs. Infomagnet generated more usability problems overall, $F(1, 22) = 18.03$, $MSE = 12.31$, $\eta_p^2 = .45$, $p < .0001$, and these problems differed in quality: Terminology, $F(1, 22) = 4.82$, $MSE = 2.50$,

Table 4. Participants' mean experience rating for the reporting methods (think aloud protocol vs. retrospective protocol) shown in bold typeface, with standard deviations in normal typeface, and *p* values for significant differences (ns = not significant). Higher means indicate more positive assessments.

	Think Aloud Protocol		Retrospective Protocol		Significance
Experience					
Difficult	**3.8**	1.1	**4.3**	1.0	ns
Unpleasant	**4.6**	0.6	**5.0**	0.0	$p = .004$
Tiring	**4.5**	0.8	**4.8**	0.4	ns
Unnatural	**4.3**	0.9	**4.1**	1.2	ns
Time-Consuming	**4.0**	1.0	**4.4**	0.8	ns
Working Conditions					
Speed	**2.5**	1.0	**3.1**	0.8	$p = .015$
Focus	**2.5**	0.9	**3.5**	0.8	$p < .001$
Presence of Experimenter					
Unpleasant	**4.5**	0.8	**5.0**	0.0	$p = .006$
Unnatural	**4.8**	1.1	**4.6**	0.6	$p = .031$
Disturbing	**5.0**	0.1	**5.0**	0.1	ns

$\eta_p^2 = .18$, $p = .03$; Data Entry, $F(1, 22) = 12.76$, $MSE = 2.04$, $\eta_p^2 = .37$, $p = .002$; Comprehension, $F(1, 22) = 40.93$, $MSE = 2.65$, $\eta_p^2 = .65$, $p < .0001$; and Navigation, $F(1, 22) = 6.18$, $MSE = .22$, $\eta_p^2 = .22$, $p = .021$. Google generated more problems of Layout, $F(1, 22) = 16.54$, $MSE = 13.94$, $\eta_p^2 = .43$, $p < .001$, and Feedback, $F(1, 22) = 4.60$, $MSE = 1.20$, $\eta_p^2 = .17$, $p = .043$. The results again seem to reflect the differing natures of the two search engines. Infomagnet is an unfamiliar tool and therefore generates more problems, whereas most users are well acquainted with Google, attuned to its conventions and less likely to question them.

3.2 Task Completion Times and Success Rates

Two within-participants ANOVAs revealed no differences in the time taken to complete the primary task for either think aloud versus retro eye cue or think aloud versus retro screen cue. The means, however, were in the predicted direction, with the think aloud condition being *slower* than either of the retrospective conditions, but the lack of reliability of this effect indicates that strong conclusions cannot be drawn concerning the reactivity of the think-aloud method. To analyse task-success a multi-dimensional chi-square test was conducted on the think-aloud condition versus the retrospective conditions. A greater number of participants in the retrospective reporting conditions completed the task (79.2%) compared to the think-aloud conditions (41.7%), which was a reliable difference, $\chi^2 = 5.58$, $df = 1$, $p = .02$.

A second chi-square test that involved a comparison between the think-aloud method and a separation of the two retrospective reporting conditions again showed a significant relationship between task completion and the protocol method used, $\chi^2 = 11.41$, $df = 2$, $p = .003$. The results indicated that it was specifically the participants in the retro screen cue condition who produced the high success rate associated with retrospective reporting, with 100% of participants within the screen capture condition completing the task successfully. In the retro eye cue condition 60% of participants completed the task, and just 42% in the think-aloud condition. The reason for the retro eye cue condition not being as successful as the retro screen cue condition may have been a consequence of a methodological error associated with the Google search engine periodically updating itself such that the sought-for answer was not actually present on the page in some trials. As the research was conducted in chunks of related conditions (albeit counterbalanced) it appeared that the retro eye cue conditions occurred more often when the answer was not on the Google results page.

3.3 Questionnaires

The questionnaire had three sections: overall experience of the method; how the method affected normal working; and the effect of the experimenter's presence. All responses were indicated on a five-point Likert scale (1 = negative assessment, 5 = positive assessment). A series of ANOVAs compared questionnaire responses in terms of the three verbalisation conditions.

3.3.1 Participant Experiences

Participants found the think-aloud method significantly more unpleasant than the retrospective methods, $F(1, 46) = 9.11$, $MSE = .21$, $\eta_p^2 = .17$, $p = .004$ (see Table 4). There were no significant differences between participants' experiences of the retrospective conditions (see Table 5).

3.3.2 Effect on Normal Working Conditions

Participants felt that they worked significantly slower and during the primary task when thinking aloud than when working silently, $F(1, 46) = 6.38$, $MSE = .78$, $\eta_p^2 = .12$, $p = .015$, and also in a less focused manner $F(1, 46) = 13.80$, $MSE = .77$, $\eta_p^2 = .23$, $p < .001$ (see Table 4).

3.3.3 Presence of the Experimenter

The presence of the experimenter had a negative effect on the participants during their think-aloud reporting, with responses indicating a significantly more unpleasant, $F(1, 46) = 8.32$, $MSE = .30$, $\eta_p^2 = .15$, $p = .006$, and unnatural, $F(1, 46) = 4.94$, $MSE = .72$ $\eta_p^2 = .10$, $p = .031$, experience than during the retrospective protocols (see Table 4). No such differences arose between the retrospective conditions (see Table 5).

Table 5. Participants' mean experience rating for the retrospective reporting methods (retro eye cue vs. retro screen cue) shown in bold typeface, with standard deviations in normal typeface, and p values for significant differences (ns = not significant). Higher means indicate more positive assessments.

	Retro Eye Cue		Retro Screen Cue		Significance
Experience					
Difficult	**4.4**	0.8	**4.3**	1.2	ns
Unpleasant	**5.0**	0.0	**5.0**	0.0	ns
Tiring	**4.8**	0.5	**4.8**	0.5	ns
Unnatural	**3.9**	1.2	**4.3**	1.2	ns
Time-Consuming	**4.5**	0.8	**4.3**	0.8	ns
Working Conditions					
Speed	**3.4**	0.9	**2.7**	0.5	$p = .03$
Focus	**3.4**	0.8	**3.5**	0.9	ns
Presence of Experimenter					
Unpleasant	**5.0**	0.0	**5.0**	0.0	ns
Unnatural	**4.5**	0.7	**4.8**	0.4	ns
Disturbing	**3.3**	0.0	**4.1**	0.1	ns

4. DISCUSSION

4.1 Usability Problems

The present study demonstrated distinct differences between the use of think-aloud protocols and the 'retro eye cue' method (i.e. retrospective protocols cued by eye tracking data). With the retro eye cue technique participants identified a greater quantity of usability problems. This finding provides some support for the hypothesis that eye-tracking data may provide a more direct contextual cue to the events of the primary task, enabling participants to retrieve "the closest approximation to the actual memory structures" [16]. In contrast, the retrospective report cued by screen replay did not promote the identification of any more usability problems than thinking aloud. The advantage of the eye cue method over the screen cue method also suggests that the difference was not due to participants having additional time and cognitive resources available to talk more freely and coherently. Consistent with previous research, which claimed that concurrent and retrospective reporting methodologies were 'equal alternatives' in usability research [18, 23, 34], we also found that the retro screen cue technique had equal standing to thinking aloud. We also note that the differences in the *type* of the usability problems identified by retro eye cue versus thinking aloud support previous findings [27, 32], and suggest that specific reporting methods may be uniquely suited to identifying particular usability issues.

Our results additionally indicate that certain retrospective reporting methods may have the ability to facilitate or inhibit the identification of usability problems, dependent on the type of search engine being tested, as evidenced by an interaction between the retrospective cue-type and the nature of the search engine being used. The novel (and arguably more complex) nature of the Infomagnet search engine particularly benefited from the eye cue method, whilst the screen cue method inhibited the number of problems detected with this search engine. In contrast, the Google search engine benefited from the screen cue method but not from the eye cue method.

4.2 Reactivity

This study also demonstrated the 'reactivity' that has previously been associated with thinking aloud [29, 30], with fewer participants successfully completing the task whilst thinking aloud than when working silently. Unlike previous studies, however, no significant difference in *task completion times* was observed across conditions. Participants rated the experience of thinking aloud negatively, indicating that it was unpleasant and adversely affected their speed of working and their task focus. These experience-based data support the view that the think-aloud method may induce negative reactivity on primary task processing as a consequence of interference and extra cognitive load [27]. The experimenter also seems to have had a more detrimental effect in the think aloud condition, with participants indicating the presence of the experimenter as more unpleasant and unnatural. The negative effect of the experimenter's presence on participants in the think-aloud condition can, perhaps, be explained by the experimenter's probing. Probing has a particularly detrimental effect during the elicitation of a think-aloud protocol because it can unintentionally lead the participant in ways that reinforce the experimenter's hypothesis. Probing can also increases the cognitive load placed on the participant in that trying to respond to the experimenter's questioning whilst completing a task is heavily demanding in terms of attentional resources.

5. Conclusions

A novel method of usability testing combining eye-tracking technology with retrospective verbal reporting has demonstrated how retrospective protocols may have value for the usability analyst if cued appropriately. As the results show, retrospective reports cued by eye-movement data can be more insightful and beneficial to the usability tester than a conventional think-aloud protocol produced concurrent to primary task processing. The possibility that retrospective protocols may show a unique suitability to certain types of search engines is an interesting finding that needs to be explored further before any firm conclusions can be drawn.

Previously identified limitations associated with the use of the think aloud method were also reproduced in the present study in terms of its potential for reactivity on primary task-based processing. In contrast, the retrospective reporting methods that we used were associated with zero interference on the primary task, as this task was completed in silence, thus maintaining the availability of cognitive resources during task completion. The

retrospective methods enabled more participants to complete the primary task successfully, and participants provided more positive statements of their experience of using these retrospective reporting methods. The analysis and coding of the retrospective reports was also facilitated by the fact that the transcripts contained more coherent sentences than those obtained using the think-aloud technique.

The general lack of appeal of the think-aloud method from the perspective of participants is interesting and seemed to be partly associated with the use of prompts during the periods when they fell silent (a finding that emerged in conversation with participants post-experiment). In contrast, the effect of prompting during the retrospective reporting methods seemed less detrimental as there were no other cognitive demands present, and the possibility of leading participants in their task behaviours was eliminated since the primary task event had already taken place. These latter benefits of retrospective reporting methods are especially important to today's usability practitioners, who often do not employ the stringent recommendations for effective elicitation of verbal reports as espoused by Ericsson and Simon [15, 16], finding them too inflexible [7]. Using a retrospective report may help control unwanted probing effects, a definite risk in the intrusive protocol techniques currently being used by some commercial usability analysts.

Despite our suggestive findings regarding the value of the retro eye cue technique as a usability-testing method, we remain acutely aware of the limitations of our research. This was a small-scale study, restricted to a single on-line search task with two search engines. To support and extend our observations future research could use the different reporting methods employed in the present study with a far wider range of search engines, browsers and websites and with a greater variety of interface tasks. Such work would help clarify which reporting methods are optimally suited to identifying particular problems with specific interaction tasks. Furthermore, in the present study the coding of reported usability problems relied on a single coder, and no attempt was made to ascribe a 'severity' rating to the problems and breakdowns identified. We are currently pursuing inter-coder reliability checks in relation to our data, and we are also systematically examining the problem-severity issue.

An additional limitation of this study is that there was no guarantee that the retrospective reports that people produced were free from post-hoc *rationalisations* of behaviours that arose during attempts to tackle the set task. In usability testing, however, the accuracy of the report produced is, arguably, not as critical an issue as it is in research that is focused on deriving a theoretical understanding the underlying nature of the cognitive processes associated with task-oriented activity. If an important usability problem is identified by means of retrospective reporting, then it is not especially relevant whether the participant truly encountered the problem during the task or whether it came to mind retrospectively. The critical point is that the potential usability problem has been identified, so that the analyst can reflect on its nature, determine its generality, and consider ways of improving the interface accordingly.

A more worrying issue, however, is the possible *omission* of real usability problems during retrospective-reporting. It is possible, for example, that the participant may understand the system far better than they originally did when examining it a second time around during the verbal-reporting stage. This could lead to the trivializing of some serious usability problems. One feature of the retro eye cue method that mitigates this possibility is that during cued retrospection the participant is not really having another attempt at the primary task, but is, instead, reflecting on what they had been doing during their previous attempt at that task. In other words, their reporting 'focus' is on their task-based behaviours rather than the task per se. We can also confirm that there was no evidence in the retrospective protocols of explicit attempts by participants to trivialise what had been fairly obvious problematic aspects of their original user experience. There was also little evidence of expressions of hesitancy or doubt about the reality of usability problems when participants produced their verbal reports. The opposite was, in fact, the case: participants were direct and clear about the problems that they had encountered; they recognized when, where and why they had experienced difficulties, and were able to explain such matters coherently. Clearly, however, it would be appealing to have a more objective method for validating the tight coupling that we believe exists between screen-based eye-movement events that are indicative of usability problems and the verbal reports of such problems that people produce retrospectively. The research reported by Ehmke and Wilson (this volume) on a correlation scheme for identifying the association between usability problems and eye-movement patterns would seem an excellent step toward having such an objective method.

In conclusion, we hope that our findings might encourage usability practitioners to examine the possible benefits that may derive from deploying appropriately-cued retrospective verbal reports in usability testing. The reduced reactivity, increased coherence, and depth of feedback that can arise from effective retrospective reporting techniques are certainly desirable aspects of any usability testing methodology.

6. ACKNOWLEDGMENTS

We thank Ursula Armitage (Bunnyfoot) and Tom Ormerod (Lancaster University) for their valuable contributions to this study. We are also grateful to Alison Walton (Bunnyfoot) for helping to pioneer the eye-movement playback technique as a way to in elicit retrospective verbal reports in usability analysis.

7. REFERENCES

[1] Ahlum-Heath, M. E., and Divesta, F. J. The effects of conscious controlled verbalizations of a cognitive strategy on transfer in problem solving. *Memory & Cognition*, *14* (1986), 281-285.

[2] Austin J., and Delaney, P. F. Protocol analysis as a tool for behavior analysis. *The Analysis of Verbal Behavior, 15* (1998), 41-56.

[3] Bainbridge, L. Verbal reports as evidence of the process operator's knowledge. *International Journal of Human-Computer Studies*, *51* (1999), 213-238.

[4] Barnum, C. M. *Usability Testing and Research*. Longman, New York, 2002.

[5] Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., and Rellinger, E. R. Metacognition and problem solving: A process-oriented approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21* (1995), 205-223.

[6] Berry, D. C., and Broadbent, D. E. Explanation and verbalization in a computer search task. *Quarterly Journal of Experimental Psychology*, *39A* (1987), 585-609.

[7] Boren, M. T., and Ramey, J. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43* (2000), 261-278.

[8] Bowers, V.A., and Snyder, H.L. Concurrent versus retrospective verbal protocols for comparing window usability. In *Proceedings of the Human factors society 34th annual meeting*. HFES Press, 1990, 1270-1274.

[9] Broadbent, D. E. Effective decisions and their verbal justification. *Philosophical Transactions of the Royal Society of London*, *B327* (1990), 493-502.

[10] Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13* (1989), 145-182.

[11] Cowen, L., Ball., L. J., and Delin, J. An eye-movement analysis of web-page usability. In *People and Computers XVI (Proceedings of HCI 2002)*. Springer, London, 2002, 317-335.

[12] Davies, S. P. Effects of concurrent verbalisation on design problem solving. *Design Studies*, *16* (1995), 102-116.

[13] Dickson, J., McLennan, J., and Omodei, M. M. Effects of concurrent verbalisation on a time-pressured dynamic decision task. *Journal of General Psychology*, *127* (2000), 217-228.

[14] Dumas, J. S., and Redish, J. C. *A Practical Guide to Usability Testing (Revised Edition).* Intellect Press, Exeter, UK, 1999.

[15] Ericsson, A. K., and Simon, H. A. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, 1984.

[16] Ericsson, K. A., and Simon, H. A. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, 1993.

[17] Gagne, R. M., and Smith Jr., E. C. A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology*, *63* (1962), 12-18.

[18] Hoc, J. M., and Leplat, J. Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies*, *18* (1983), 283-306.

[19] Jacob, R. J. K., and Karn, K. S. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyona, R. Radach, and H. Deubel (eds.), *The mind's eye: Cognitive and applied aspects of eye movements.* Elsevier London, UK, 2003, 573-605.

[20] Kirk, P., and Ashcraft, M. H. Telling stories: The perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27* (2001), 157-175.

[21] McCarthy, J. D., Sasse, M. A., and Riegelsberger, J. "Could I have the menu please?": An eyetracking study of design conventions. In *People and Computers XVII (Proceedings of HCI 2003).* Springer, London, 2003, 401-414.

[22] Nisbett, R. E., and Wilson, T. D. Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84* (1977), 231-259.

[23] Nielsen, J. *Usability Engineering*. Academic Press, Boston, MA, 1993.

[24] Nielsen, J., Clemmensen, T., and Yssing, C. Getting access to what goes on in people's heads: Reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction (NordiCHI 2002)*. ACM Press, New York, NY, 2002, 101-110.

[25] Poole, A., and Ball, L. J. Eye tracking in HCI and usability research. In C. Ghaoui (ed.), *Encyclopedia of human-computer interaction.* Idea Group Inc., Pennsylvania, 2006.

[26] Poole, A., Ball, L. J., and Phillips, P. In search of salience: A response time and eye movement analysis of bookmark recognition. In *People and Computers XVIII (Proceedings of HCI 2004)*. Springer, London, 2004, 363-378.

[27] Robinson, K. M. The validity of verbal reports in children's subtraction. *Journal of Educational Psychology*, *93* (2001), 211-222.

[28] Rubin, J. *Handbook of Usability Testing: How to Plan Design and Conduct Effective Tests*. Wiley, New York, 1994.

[29] Russo, J. E., Johnson, E. J. and Stephens, D. L. The validity of verbal protocols. *Memory & Cognition*, *17* (1989), 759-769.

[30] Schooler, J. W., Ohlsson, S., and Brooks, K. Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122* (1993), 166-183.

[31] Schuck, J. R., and Leahy, W. R. A comparison of verbal non-verbal reports of fragmenting visual images. *Perception and Psychophysics*, *1* (1966), 191-192.

[32] Taylor, K.L., and Dionne, J. P. Accessing problem solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *29* (2000), 413-425.

[33] Teague, R., De Jesus, K., and Nunes-Ueno, M. Concurrent vs. post task usability test ratings. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 2001).* ACM Press, New York, NY, 2001, 289-290.

[34] Van den Haak, M. J., de Jong, M. D. T., and Jan Schellens, P. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22* (2003), 339-351.