An Eye Movement Analysis of Webpage Usability

Laura Cowen†‡, Linden J. Ball† and Judy Delin‡

†Department of Psychology, Lancaster University, Lancaster LA1 4YF, UK
‡Enterprise IDU, Old Chantry Court, 79 High Street, Newport Pagnell, Buckinghamshire, MK16 8AB, UK

Tel: *+44 01524 593470*
Fax: *+44 01525 593744*
Email: *l.ball@lancaster.ac.uk*

An experiment is reported that investigated the application of eye movement analysis in the evaluation of webpage usability. Participants completed two tasks on each of four website homepages. Eye movements and performance data (Response Scores and Task Completion Times) were recorded. Analyses of performance data provided reliable evidence for a variety of Page and Task effects, including a Page by Task interaction. Four eye movement measures (Average Fixation Duration, Number of Fixations, Spatial Density of Fixations, and Total Fixation Duration) were also analysed statistically, and were found to be sensitive to similar patterns of difference between Pages and Tasks that were evident in the performance data, including the Page by Task interaction. However, this interaction failed to emerge as a significant effect (although the main effects of Page and Task did). We discuss possible reasons for the non-significance of the interaction, and propose that for eye movement analysis to be maximally useful in interface-evaluation studies, the method needs to be refined to accommodate the temporal and dynamic aspects of interface use, such as the *stage* of task processing that is being engaged in.

Keywords: eye movement analysis, webpage usability, performance measures.

1 Introduction

There is currently a multitude of methods available for evaluating user interfaces. These range from subjective user feedback (including interviews and focus groups), through semi-formal methods such as Cognitive Walkthroughs (Wharton et al, 1994; Preece et al, 1998) and Heuristic Evaluation (Nielsen, 1994), to more objective user testing. Although the latter is probably the most reliable evaluation technique, it remains rather limited in the amount of information it can provide about user performance. Typically the only data that are acquired via user testing are success rates and completion times for users attempting interface tasks. These data can inform designers about when the user had difficulties with the interface, but not necessarily what specific areas of the interface caused such problems.

The interest in finding objective, usability-evaluation methods that can pinpoint problematic features of interfaces has prompted researchers to look at how eye movements might be used to understand the way that users view, search and process interface information (e.g. Baccino & Colombi, 2001; Crowe & Narayanan, 2000). The present paper reports an exploratory experiment investigating the use of eye movement measures to evaluate webpage designs. We argue that eye movement data can augment the data obtained through user testing by providing more specific information about the user's cognitive processes (see Rayner, 1995, and Henderson & Hollingworth, 1999, for discussions of the relationship between eye movements and cognition).

1.1 What Makes Eye Movement Analysis Useful?

When an individual looks at an object, an image of the object is projected on to the retina, which is composed of light-sensitive cells that convert light into signals that can be transmitted to the brain via the optic nerve. The distribution of these retinal cells is uneven, with denser clustering at the centre of the retina than at the periphery. Such clustering

causes the acuity of vision to vary, with the most detailed vision available when the object of interest falls on the centre of the retina. Outside this foveal region visual acuity rapidly decreases. Eye movements are made to reorient the eye so that the object of interest falls upon the fovea and the highest level of detail can be extracted (Gregory, 1990; Rayner & Pollatsek, 1992).

The focusing of the eye on an object is termed fixation. A fixation typically lasts about 300ms. After a fixation the eye goes through a movement—termed a 'saccade'—to fixate on another part of the same object or on a new object. Such saccades are high-speed, ballistic movements that last approximately 150-200ms from planning to execution (Palmer, 1999). During a saccade no information is obtained as perception is inhibited to prevent the viewer seeing a blur. Only when the eye is relatively still, during a fixation, can information be extracted from the display. The assumption of researchers who aim to examine eye movements in order to assess the usability of displays and computer interfaces is that the duration of fixations and the pattern of eye movements in general are dependent on how easy or difficult the display is to process. If the display, or any part of it, is difficult to process then fixations will be longer and there will be more fixations closer together (with relatively short saccades) than if the display is easy to process.

1.2 Using Eye Movements to Evaluate Interface Usability

A major problem with using eye movements to evaluate interface usability issues is that eye movement recordings provide a large quantity of raw data that can be time-consuming to analyse and potentially difficult to interpret meaningfully. In their pioneering research on eye movement analyses of interface usability, Goldberg and Kotval (1998, 1999, in review; Kotval & Goldberg, 1998) proposed a set of 11 eye movement measures that they argue can make analysis more efficient and can also be automated to some extent (see Goldberg and Kotval, 1989, for full details). Included in these are *spatial measures*, such as the number of backtracking saccades made on the display, which may indicate whether or not an interface matches up with a user's expectations (i.e. if a user's expectations are fulfilled then they shouldn't have to move their eyes back and forth several times across the display). Other examples of spatial measures include the total scanpath length (i.e. the total distance the eyes move around the interface), the spatial density (or distribution) of fixations on the interface, and the average saccade length or amplitude (which would indicate the extent of search, and, therefore, the quality of the layout). Goldberg and Kotval also propose a range of *temporal measures* that indicate the depth of processing required by an interface user. These include measures of the mean duration of fixations while using the interface, and the ratio of fixation to saccade duration (which indicates the relative proportions of time spent processing and searching the interface).

Goldberg and Kotval evaluated the validity of these measures by examining user's eye movement behaviour in relation to several versions of a Windows-style interface that they created in order to simulate a graphical software package. Down one side of the interface was a selection of buttons, or 'tools', that the participant was required to find whilst their eye movements were recorded. The physical grouping or appearance of the tools was varied in different experimental conditions so as to make the tasks harder or easier to perform. The difficulty of the tool groupings was assessed independently by 80 interface design experts and typical users. The investigators examined whether the proposed eye movement measures would be sensitive to the differences between the interfaces, and also whether the findings produced by the eye movement measures positively correlated with the independent usability ratings of the interfaces.

It was found that the various measures were differentially sensitive to manipulations of the experimental interface layout. So, for instance, when the physical grouping of the tools on the interface was varied, only measures of global searching behaviour tended to be sensitive because little processing was required of the tools themselves once found. However, when the tools varied in the way in which they were labelled, nearly all of the measures were seen to be sensitive to such differences, since different levels of search and processing behaviour were required by the alternative interfaces.

When the data from the measures were correlated with the independent usability ratings, all the measures indicating global search behaviour were found to have linear relationships with the usability ratings, with the best predictors of usability being the number of backtracks and scanpath duration. Indeed, an unusable interface could be predicted to necessitate

scanpaths nearly 80% longer than those required by an excellent interface. Local search and processing measures (at the level of individual fixations and saccades) showed interesting patterns of sensitivity to usability. The relationships between the fixation duration and fixation/saccade ratio measures and usability were both U-shaped. Therefore, at the lowest and highest levels of usability, the measures were largest, decreasing at intermediate levels of usability. However all three local search and processing measures showed less than 50% change between the best and poorest interfaces, and, therefore, were not as sensitive to variation in usability as some of the global search measures.

1.3 The Use of 'Real' Stimuli in Usability Evaluation

Although Goldberg and Kotval's work motivated aspects of our present study, the experimental design that we adopted differed from theirs in several important ways. Perhaps most crucial of all were the differences between the two experiments in terms of the nature of the interfaces that were used as stimuli and the tasks that participants were required to undertake.

Looking at the stimulus issue first of all, Goldberg and Kotval created their interface designs especially for use in their experiments. Although this method of stimulus design allows for a large amount of control over variables, it is difficult to generalise findings to real interfaces designed by real designers for real end-users. Most interfaces used by people in everyday contexts have been designed for them to complete tasks that satisfy their personal goals. In the present experiment, home pages from commercial websites were used as stimuli, as web design is currently of great interest to usability specialists.

Concentrating on web interfaces in the present study made it relatively easy to obtain a selection of complementary, but real webpages. An alternative approach to obtaining a selection of 'naturalistic' webpage designs might have been to take a single page design and to adapt it multiple times to produce different stimuli. However, as argued by Buckingham (1931) on the subject of manipulating typographical layouts for experimental purposes, "[the]…separating of size of type, length of line, and interlinear spacing…is wholly artificial." (p.103). By this he meant that it is not possible, in testing visual designs, to manipulate only one variable whilst keeping all others constant.

For example, in the present situation, moving a webpage's navigation menu to different locations on the page gives one variable that can be manipulated easily: menu location. However, it is likely that other, less obvious variables, are also created when the navigation menu is moved, for example, the amount of empty space increases where the menu used to be, and the empty space is reduced in the menu's new location. Also, it may be necessary to move other items on the page in order to accommodate the new position of the navigation menu, which would create yet another variable.

The webpages selected for the present study were homepages from four websites produced by an international mobile phone service provider. In each country that the company operates, there is a separate website. Only English-language websites were used: from Belgium, India, Switzerland and the UK. The websites were appropriate for experimental testing because they each had an individual design (page layout, content, information architecture), whilst holding constant the overall 'look' of the company's brand (e.g. the standard colour scheme, the presence of the corporate logo on each homepage, and similar subject matter and purpose). Therefore, these websites naturally possessed both variables that were held constant and variables that varied from stimulus to stimulus.

A second critical difference between our study and that of Goldberg and Kotval concerned the interface tasks that participants were asked to attempt. Goldberg and Kotval's participants were required simply to find, as quickly as possible, a tool on the interface. This, however, appears to be a rather simplistic and de-contextualised interface activity. Also, since the tasks were presented in the middle of the interface for the participants to read, the experiment would have taken away most of the realism of the exercise. In the present study the interface tasks were read to the participants and consisted of requests to find pieces of information of a kind that would be required by a real user of the website.

Another major difference between the two studies was in the way in which the eye movement measures were evaluated. Goldberg and Kotval had 50 typical users and 30 interface designers rate the interfaces for their usability and then correlated such ratings with eye movement scores. However, it is debatable whether such usability ratings are themselves valid. The 30 interface designers may well have had experience of (and possibly training in)

predicting the usability of an interface, but the 50 typical users were unlikely to have had such experience or training.

In the present experiment the relative usability of the four stimulus pages was evaluated by analysing the participants' task performance independently from their eye movement behaviour. This meant that the results of the eye movement measures could then be compared with the findings from the performance measures in order to assess whether they detected the same pattern of differences between pages.

It is finally noteworthy that fewer eye movement measures were used in our study than in Goldberg and Kotval's study. Three of their measures were employed: Average Fixation Duration (a processing measure), Number of Fixations (a local search and processing measure), and Spatial Density (a global search measure). A fourth measure that they did not use—Total Fixation Duration—was employed as a global measure of the total amount of processing performed on each page, rather than just the mean amount of processing on each part of a page.

2 Overview of the Experimental Design and Predictions

Each of the four webpages that we used in the experiment was presented twice to each participant. Appropriate counter-balancing was used to counteract potential order effects. On viewing a webpage, the participant was asked to find information about either using a mobile phone abroad (Task 1) or buying a new mobile phone handset (Task 2). All participants completed both tasks on all the pages.

Three hypotheses were derived for the study in relation to participants' performance data, that is, their task processing time and success rates. First, it was predicted that the four pages would differ in their support of the participants' tasks (i.e. there would be a main effect of the *Page* factor). Second, it was predicted that Task 1 ('abroad') would be seen to be more difficult than Task 2 ('handset'), since the concept of how to go about using a mobile phone abroad would be less familiar to participants than the concept of shopping (i.e. there would be a main effect of the *Task* factor). Third, the relative difficulty of the two tasks would depend on the pages they were being performed on (i.e., an interaction would be evident between the Page and Task factors). Finally, our expectation was that some—and possibly all—of the eye movement measures should detect the differences described in the previous three hypotheses.

3 Method

3.1 Participants

Seventeen participants took part in this experiment. Their ages were classified within five-years intervals. The modal interval was 25-29, with seven of the participants within this age range. The other participants were spread across the intervals tailing off at the ends: 15-19 and 45-49. All participants either did not wear glasses or did not need them to read the displays used in the study. Approximately half the participants had taken part in eye tracking experiments previously. The participants came from a variety of professions and included students, graphic designers, writers, teachers and lecturers.

All participants were regular web users and all but one owned a mobile phone. English was the native language for all but two participants, and these were currently completing doctoral degrees at UK universities, having previously also studied in the UK and, therefore, they spoke and read English to a high standard. Participants took part in the study voluntarily and were typically interested in the eye tracker and wanted to watch the video of their own performance after completing the experiment.

3.2 Apparatus and Stimuli

Three PCs running MS Windows 98 were used to run the experiment and record data.

3.2.1 Stimulus Presentation

Stimuli were presented on a Pentium II 400Hz desktop PC with 128Mb RAM and 16Mb display memory. The monitor had a 15 inch flat LCD screen (for clarity of external video

recording) with a screen area of 1024x768 pixels. The monitor was placed on a stand to raise it up so that the centre of the display was level with the participant's eyes. The keyboard was not required for the experiment and was hidden. Participants made their responses by clicking on hyperlinks with a mouse.

3.2.2 Data Collection

Eye movements were recorded using SMI's Head-mounted Eyetracking Device II (HED-II) with Scene Camera. The eye tracker uses two small cameras (the Eye Camera and the Scene Camera) mounted on a bicycle helmet for comfort, weighing only 450g in total. No contact is made with the participant's eye. A harmless and unnoticeable infrared light shines into the participant's eye so that the front surface of the eyeball is illuminated. This produces two effects: the bright pupil and the corneal reflection (SensoriMotoric Instruments, 1999). Because the eyeball is not a perfect sphere, the corneal reflection moves less than the pupil as the eye rotates. The image-processing software, 'iView', on the Pentium II PC (to which the eye tracker is connected), analyses the transmitted image. It computes the centres of the pupil and corneal reflection, from which it can compute the Point of Regard (SensoriMotoric Instruments, 1999) or absolute line of gaze (Jacob, 1995). The resolution of the HED-II eye tracker is better than one degree and it has a sampling rate of 50Hz. Every 20ms the eye tracker transmits the x and y coordinates of the participant's visual line of gaze to the iView PC for processing.

The eye movement data collected in the present study, including the x and y coordinates of gaze and also the processed fixations, were saved for subsequent analysis, and exported to spreadsheets. Recording of the eye movement data files had to be started and stopped manually by the experimenter. In addition, the video recording from the Scene Camera of the display the participant was viewing was also saved. A still snapshot was taken from the video so that the eye movement scanpaths could be overlaid for calculating data for the eye movement-derived Spatial Density measure. Before the experimental trials began, each participant was calibrated to the screen of the monitor on which the stimuli were to be presented. A white sheet of paper was taped over the screen, to prevent distraction by items displayed on the computer desktop. A laser pointer was used to mark nine points on the paper for the participant to fixate in turn. Participants were seated on a stable chair and a chin-rest was provided so that they could hold their head steady during the testing.

3.2.3 Stimuli

The four website homepages described in Section 1.3 were used as stimuli. Two tasks were asked of each participant on each page. Task 1 stated that: "You want to know more about using your mobile phone outside <country name>", whilst Task 2 stated that "You want to buy a new mobile phone *handset* from this website". The tasks were worded in the second person to try to encourage the participant to think of the task as their own goal. The name of the country (in which the website was used) was used in Task 1 rather than saying 'abroad', which could have primed participants to look for a particular term on the page. In Task 2, the word 'handset' was emphasised so that participants understood that they were looking for phones rather than whole packages and payment plans.

So as to avoid participants becoming too familiar with the pages, the order of task presentation was counterbalanced. All participants viewed the pages (labelled A to D) in the same order: A, D, B, C, D, A, C, B. Participants were allocated alternately to group 'Order 1' or group 'Order 2'. The Order 1 group received Task 1 on the first page (A), Task 2 on the second page (D), Task 1 on the third page (B), and so on. The Order 2 group received Task 2 on Page A, Task 1 on Page D, and so on. Each participant received each task once on each page but never received the same task on two consecutive pages and never received the same page without two other pages in-between.

Pages were loaded prior to participants entering the room. The windows were maximised to fill the full screen and then retracted to the Windows Task Bar so that they could not be seen by the participant prior to starting the experiment. The pages were presented individually in Internet Explorer 5.5 windows so that all the animations and links worked correctly. The presentation of pages was controlled by the participant. The participant opened the correct page when asked to do so by the experimenter. The tasks were read aloud to the participant who was asked to avoid head movements (including speaking) once calibrated.

3.3 Procedure

It was explained to participants that four homepages from mobile phone international websites would be presented twice each and that for each page there would be a task to complete. They were told to complete the task by clicking on the link where the information could be found. If they had difficulty with the task they were to guess at which was the correct link. The procedure for opening and closing pages using the computer mouse was explained until the participant indicated understanding of this. The participant was shown the eye tracker and the experimenter explained what the cameras did. It was emphasised that only the eye and the stimulus display would be recorded, not the participants themselves.

After gaining consent to continue the study, the tracker helmet was placed on the participant's head and fastened, and the participant's eyes were calibrated to the display screen. Once calibrated, the participant was asked not to move, particularly during the presentation of a page. In between pages, if necessary, they could move slightly to get more comfortable. All participants were offered the opportunity to sit back for a moment half-way through the experiment.

The experimenter started the recording of the dot-overlaid scene video on the laptop and this recorded continuously until the end of the testing session. The participant was asked to open the first page. As soon as the page appeared on-screen, the experimenter started recording the eye movement data using the iView software on the iView PC (this had to be done manually by pressing a key: once to start recording and once to stop), while simultaneously starting to read aloud the relevant task. The participant could look at the page and use the mouse at any time during the presentation of the page stimulus (including while the experimenter was reading the task) and the trial was completed when they clicked on a hyperlink, whether or not it was the correct target. When the participant clicked a hyperlink, the experimenter stopped iView recording the eye movement data. At the end of the testing session, after the participant had answered each task on each of the four pages, the dot-overlaid scene video recording was stopped. The Mpeg produced by the dot-overlaid scene video recording was saved, as was the eye movement data file and the associated fixations file.

4 Results and Discussion

Two classes of data were obtained: performance data and eye-movement data. The performance data taken were Response Scores (whether or not a correct link was clicked), and Task Completion Times (how long it took the participant to complete the task—and self-terminate the trial—by clicking a link). The eye-movement data involved four different measures that are described in detail in Section 4.2. With the exception of Response Scores, all data were examined statistically using analysis of variance (ANOVA). The factors in these ANOVAs were Page (within participants, with four levels: A, B, C, and D), and Task (within participants, with two levels: Task 1—'abroad' vs. Task 2—'handset'). Where necessary, Simple Main Effects analyses and Tukey HSD tests were also conducted. The alpha level for all tests was set to .05. The data for each measure, apart from Response Scores, were tested for skew prior to analysis. As some conditions produced skew values greater than 1, the data were log-transformed. For some measures this reduced the levels of skew, but when the transformed data were subjected to ANOVA the effects revealed were similar to those obtained when analysing the untransformed data. All results that we present in this paper are, therefore, based on the untransformed data.

4.1 Performance Measures

4.1.1 Response Scores

Participants' performance was scored for each task on each page. If the link that was clicked (terminating the trial) would have provided the information required by the task then a score of 2 was given. If the link would not have provided this information then a score of 1 was given. Table 1 presents the percentage frequency of correct responding for each task on each page. This table reveals that participants performed poorly on page D and extremely well on

Page B, whilst Pages A and C were intermediate. Overall, there appears to be no major effect of Task on responding, although there is some indication that Task 1 (‘abroad’) was slightly harder than Task 2 (‘handset’). There is also a suggestion that the Page and Task factors may interact as there is a marked separation between performance levels for Tasks 1 and 2 on Page A that is not evident for other pages.

As the Response Scores represent a dichotomous dependent variable it was inappropriate to subject these data to ANOVA. Instead, the Page and Task factors were treated as predictor variables, and logistic regression was used to examine whether they were predictive of Response Score. All predictors (including the Page x Task interaction) were entered into the regression model in a single block. The variability explained by the model was good (41%—Cox & Snell R Square; 56%—Nagelkerke R Square) and the overall percentage of responses correctly predicted by the model was 80%. This represented a marked improvement on the 65% value based on the initial maximum likelihood estimation. The overall model was significant at $p = .001$ according to the Chi-square statistic (67.57, $df = 7$). The regression analysis indicated that only the effect of Page was significant (Wald statistic = 9.70, $p = .021$, $df = 3$), demonstrating the high predictive validity of this variable as a determinant of correct responding. In sum, the logistic regression analysis confirmed the effect of Page on Response Scores that is evident in the descriptive data presented in Table 1, albeit with the caveat that that the apparent Page x Task interaction was not reliable on this performance measure.

Table 1: Percentage frequency of ‘correct’ responding broken down by Page and by Task.

	Page A	Page B	Page C	Page D	
Task 1	35	100	65	38	60
Task 2	100	100	65	13	70
	68	100	65	26	

Inspection of the percentage frequency correct scores in Table 1 shows that only on Page B did all of participants correctly answer both tasks. However, as Table 2 shows, the four pages varied in the number of correct target links available. As can be seen, Page B was also the page with the most correct targets available, three for each task, suggesting that the probability of clicking the correct link depended on how many correct links were available. However, although Page A also contained three correct possible targets for Task 1 (‘abroad’), the frequency of correct responding here was the lowest for that Task across the four pages. Notwithstanding the fact that this difference was not reliable in the logistic regression analysis, it does hint at the possibility that the design of a page may well ‘hide’ the correct responses from the user, leading to lower levels of performance. Future work could systematically manipulate the number of correct target links associated with tasks in order to examine effects on performance.

Table 2: Number of correct target links available on each page.

	Page A	Page B	Page C	Page D
Task 1	3	3	2	2
Task 2	2	3	1	1

4.1.2 Task Completion Times

The duration of time between the page appearing on-screen and the participant clicking on a link in response to the task (i.e. task completion) was recorded. Each participant could provide a maximum of eight task times (one for each task on each page). Eleven data points were removed prior to analysis due to confounds such as the participant mishearing the task and asking for it to be repeated, or pages loading slowly or incorrectly. The mean task completion times on each page are presented in Table 3.

A 2 x 4 ANOVA revealed significant main effects of Page [$F(3,24) = 11.28$, $p < .001$], and Task [$F(1,8) = 11.29$, $p = .010$], as well as a Page x Task interaction [$F(3,24) = 5.57$, $p = .005$]. Simple main effects analyses were conducted to understand the cause of the observed interaction. It was found that on Page A—but not on the other pages—performance at Task 1 was significantly slower than at Task 2 [$F(1,14) = 21.23$, $p < .001$]. This suggests that on Page A, finding out about buying a new handset was better supported than finding out about taking a phone abroad. An effect of Page was found for Task 1 ('abroad') [$F(3,33) = 23.11$, $p < .001$] and further analyses using the Tukey test showed that performance was reliably slower on Page A than on all the other three pages. An effect of Page was also found for Task 2 ('handset') [$F(3,33) = 5.06$, $p = .005$] and was also further analysed using the Tukey test to show that performance was reliably slower on Page D than on Page B and Page C. Finding out about using a phone abroad was more difficult on Page A than on any other but finding out about new handsets proved most difficult on Page D.

These various Task Completion Time results were similar to the pattern of results relating to the Response Scores measure (Table 1), although this latter performance measure does not appear to have been sensitive enough to detect reliable effects for all factors. On the whole, though, it is evident that pages that produced more correct responses also produced the faster response times. Responses on Page A were more often correct on Task 2 ('handset') than on Task 1 ('abroad'). Furthermore, these responses were given more quickly on Task 2 than on Task 1, suggesting that participants had more trouble completing Task 1 than Task 2, even though there was one more correct link available for the former task than for the latter (Table 2) on Page A. The observations that performance for Task 2 was significantly slower on Page D than the other pages, and that performance of Task 1 was significantly slower on Page A than the other pages, also corroborate similar findings revealed by the Response Scores measure. Taken together, these similarities in the findings of the two measures provide a solid benchmark of the relative levels of usability of the four pages. The findings of the eye movement measures can, therefore, be compared confidently with the findings relating to these performance measures.

Table 3: Mean task completion time (s).

	Page A	Page B	Page C	Page D	
Task 1	28.112	13.881	12.623	15.825	17.610
Task 2	14.975	12.269	11.415	21.277	14.984
	21.544	13.075	12.019	18.551	

4.2 Eye Movement Measures

The measures taken of the recorded eye movement scanpaths were: *Total Fixation Duration* (the sum of all the fixation duration times whilst completing a task); *Number of Fixations* (the total number of individual fixations on a page whilst completing a task); *Average Fixation Duration* (the average duration of individual fixations on a page whilst completing a task); and *Spatial Density of Fixations* (the spatial distribution of fixations on the page whilst completing a task). Eye movements were collected at a sampling rate of 50Hz (the x and y coordinates of the eye's line of gaze were collected every 20ms). These gazepoint samples were processed by the iView software to form fixations and saccades. The minimum duration of a fixation was 100ms and the gazepoint could not deviate more than 40pts horizontally or vertically during a fixation[1]. Each participant could provide a maximum of eight scores for each measure, one for each of two tasks on each of four pages. Due to one participant being poorly calibrated, only 16 of the 17 participants could provide eye movement data for analysis. In addition, 15 data points (16 on the Spatial Density measure) were removed from each measure prior to analysis due to technological problems, head movement or poor calibration.

[1] "Gaze position in iView has no physical unit (e.g. millimeters, pixels) but is related to the calibration area settings." (SMI, 1999). The calibration area in the present experiment was 721pts x 279pts.

4.2.1 Total Fixation Duration

The duration of all the individual fixations made on a page during the performance of a task were totalled. The mean total fixation duration can be found in Table 4. A 2 x 4 ANOVA was conducted and revealed main effects of Page [$F(3,12) = 5.62$, $p = .012$], and Task [$F(1,4) = 7.93$, $p = .048$]. The main effect of Task revealed that participants spent more time fixating (processing) when completing Task 1 (‘abroad’) than when completing Task 2 (‘handset’). This suggests that Task 1 was probably more difficult to represent mentally than Task 2. A significant main effect of Task was also found in the Task Completion Times measure reported above—though in the latter it was also shown that the effect of Task interacted with the effect of Page. Within this interaction, where there was found to be a significant simple main effect of Task (though on Page A only), it was (as here) Task 1 that was found more difficult than Task 2.

The simple main effect of Page was further analysed using the Tukey test. This revealed that total fixation duration (i.e. total processing time) on Page A was significantly longer than on both Page B and Page C. The Total Fixation Duration measure is linked with the Task Completion Time measure but is not the same because the latter includes the duration of saccades as well as fixations. The Total Fixation Duration measure produced similar but not identical results to the Task Completion Time measure. It only detected differences in processing times rather than differences in overall performance (processing and searching) times.

Table 4: Mean total fixation duration (s).

	Page A	Page B	Page C	Page D	
Task 1	23.502	14.880	14.937	14.904	17.056
Task 2	14.674	12.053	9.829	17.440	13.499
	19.088	13.467	12.383	16.172	

4.2.2 Number of Fixations

The number of individual fixations made on a page whilst performing a task was totalled. As the 2 x 4 ANOVA that was conducted on these data revealed no significant effects, the mean values for this measure have not be tabulated here.

4.2.3 Average Fixation Duration

For each task on each page, the total fixation duration was divided by the number of individual fixations to reveal the average duration (in milliseconds) of fixations on a page whilst completing a task. The mean fixation duration can be found in Table 5. A 2 x 4 ANOVA revealed a single significant main effect of Page [$F(3,12) = 5.45$, $p = .013$]. Further analysis using the Tukey test showed that this effect was due to the individual fixation durations on Page C being significantly shorter than on Page A. This implies that Page A had a lower level of usability than Page C. Inspection of the means for each of the other measures for the main effect of Page showed Page A to consistently be the one with the lowest usability (the highest mean), except on the Response Scores measure where Page A had the second highest overall response score. However, the interaction revealed by the Task Completion Time performance measure showed that the usability of Page A depended significantly on which task was being performed on it. So although the Average Fixation Duration measure was sensitive to the general overall pattern of page means, it was not sensitive enough to detect the more complex Page x Task interaction in determining a page’s level of usability.

Table 5: Mean fixation durations on each page (ms).

	Page A	Page B	Page C	Page D

Task 1	553.50	488.00	404.83	434.83	470.29
Task 2	487.50	471.33	394.00	482.50	458.83
	520.50	479.67	399.42	458.67	

4.2.4 Spatial Density of Fixations

The scanpaths of the recorded eye movements were overlaid (using the iView software) on snapshots of the display taken from the Scene Camera video. A screenshot was taken and loaded into a graphics package so that a grid, with squares measuring 30 x 30 pixels, could be overlaid on the screenshot of the page. The number of squares containing one or more fixations was manually counted. This number was divided by the total number of squares covering the page and then multiplied by 100 to produce a percentage. A 2 x 4 ANOVA revealed no significant differences in spatial coverage of fixations regardless of Task or Page, and so mean values for this measure have not be tabulated here.

4.3 Relative Usability of the Four Homepages

The performance measures (Response Scores and Task Completion Times) were collected to identify differences in usability (how well a page design supported the user's task) between the four pages. The more sensitive Task Completion Time measure revealed a significant interaction between Page and Task, indicating that the usability of a page depended on the task performed on it. Further investigation into the nature of this interaction showed that Page A supported the task of finding out about new handsets significantly better than the task of finding out about taking a phone abroad. This basic difference of task difficulty on Page A was reflected in an omnibus main effect of Task for the Task Completion Time measure, where the task requiring information about taking a phone abroad was significantly more difficult than the task requiring information about new handsets.

Although many of the differences in usability between individual pages were not significant, the two measures showed similar patterns of results (Figures 1 & 2; Tables 1 & 3). On the whole, Page A and Page D showed the worst usability (despite the good performances on Task 2 on Page A) while Page B and Page C showed intermediate to good usability. Most pronounced on the graphs is the difference in the performances for the two tasks on Page A. As a high score shows good usability on the Response Scores measure but bad usability on the Task Completion Times measure, the two graphs show very similar patterns of results. Therefore, it is probably safe to compare the relative sensitivity of the eye movement measures with that of the performance measures.

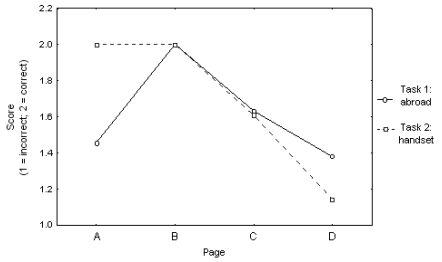

Figure 1: Mean Response Scores (where 2 = correct response, and 1 = incorrect response).

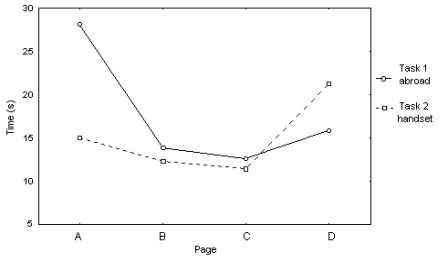

Figure 2: Mean Task Completion Times: Page x Task interaction [F(3,24) = 5.57; p = .005].

4.4 The Sensitivity of Eye Movement Measures to Usability

The eye movement measures (Total Fixation Duration, Number of Fixations, Average Fixation Duration, Spatial Density) were collected in order to observe whether they were sensitive to the relative usability differences revealed by the two performance measures. Unlike the Task Completion Time performance measure, none of the eye movement measures was sensitive to the Page x Task interaction. The two time-based fixation measures did, however, detect

significant main effects of Page in line with both performance measures. Total Fixation Duration showed that Page A required significantly more processing overall than both Page B and Page C; the Average Fixation Duration measure showed that Page A required significantly more local processing than Page C. A main effect of Task from the Total Fixation Duration measure did show a difference in the same direction as that found by the Task Completion Time measure: the 'handset' task was found easier than the 'abroad' task. These results from the two time-based eye movement measures, along with the similar patterns of page usability differences (Figures 1-4) suggest that the four measures (two performance and two time-based fixation measures) were detecting similar variations in usability across the four pages.

The other two measures, though not showing significant differences for either main effects of Page or Task, or for Page x Task interactions, do produce graphs (Figures 5 and 6) that are similar in shape to those produced by the results from the other measures. The eye movement measures do appear to have been able to detect the same usability differences between the pages as the performance measures. Reasons why the eye movement measures were not as reliably sensitive as the performance measures are considered in the General Discussion.

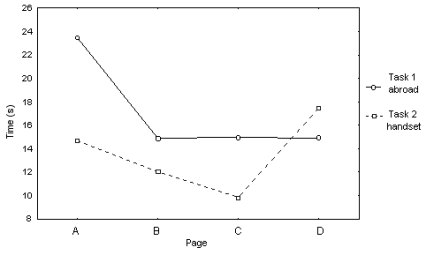

Figure 3: Mean Total Fixation Duration: Page x Task interaction [F(3,12) = 1.00; p = .426].

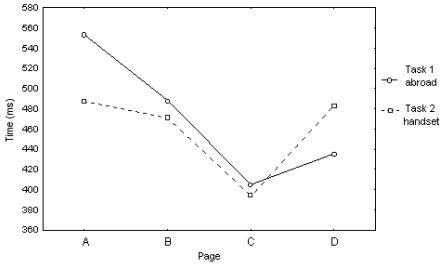

Figure 4: Mean Average Fixation Durations: Page x Task interaction [F(3,12) = 1.41; p = .287]

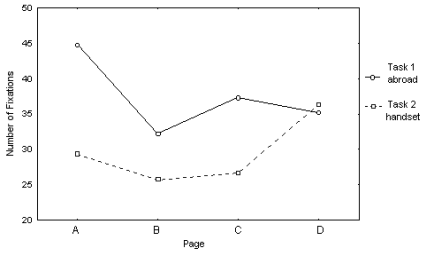

Figure 5: Mean Number of Fixations: Page x Page x Task interaction [F(3,12) = 0.41; p = .746].

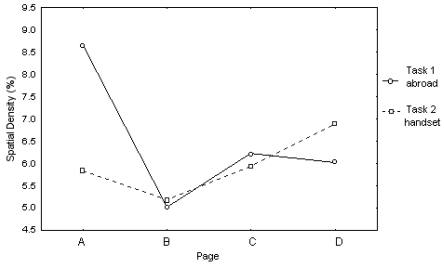

Figure 6: Mean Spatial Densities: Page x Task interaction [F(3,9) = 0.75; p = .548].

5 General Discussion

A number of a priori hypotheses were tested in this exploratory study. The first prediction was that the four pages would differ in their support of the given tasks. Significant main effects of Page were found in the data of both performance measures and also the two time-based eye movement measures. The second prediction was that Task 1 would be seen to be more difficult than Task 2. Significant main effects of Task were found in the data of one performance measure (Task Completion Time) and one eye movement measure (Total Fixation Duration). Our third prediction was that the relative difficulty of the two tasks would depend on the pages they were being performed on. A significant interaction between Page and Task was found in the data for the Task Completion Time performance measure. None of the eye movement measures found this interaction to be significant, but inspection of the graphs of means for each measure reveal patterns concomitant with the emergence of a Page x Task interaction effect.

Overall, our findings lend support to the view that the eye movement measures we took were sensitive to similar patterns of influence that were evidenced in the performance measures that we derived. However, some of the key differences found to be significant in the performance data did not emerge as significant effects in the eye movement data. It is possible, however, that the eye movement measures were sensitive to more than one factor

influencing a page's usability, and maybe two or more factors were cancelling out each other —so reducing the sensitivity of a measure. In this respect it is noteworthy that the eye movement measures of local search and processing in Goldberg and Kotval's (in review) experiment were found to show quadratic relationships with the usability ratings. Goldberg and Kotval concluded that that the longer fixations and shorter saccades at both higher and lower levels of usability were not necessarily due to the same factor. It is likely that there is more than one type of search behaviour and more than one type of processing behaviour exhibited by users at an interface. Interpreting long fixations as being due to extended processing, and large spatial densities as being due to inefficient searching, is rather vague.

If different search strategies and processing types could be identified, and the eye movement patterns associated with them, then it might be possible to break down measures such as average fixation duration into separate measures according to search strategy or stage of processing. For example, the average fixation duration of all instances of stage 'a' processing could be calculated. All instances of stage 'b' eye movement patterns would be calculated separately, as would instances of stage 'c'. The interface could then be more precisely evaluated for which stages of processing it fails to support, rather than just saying it fails to support any type of processing. Assuming that several processing stages do exist and that the eye movement measure is capturing them all, then if one stage is supported by the interface but not the others, any effect is likely to be cancelled out. The eye movement measure used, therefore, would appear to be insensitive to all processing when, in fact, it might be trying to detect three different types of processing.

Although many differences were not significant in the data from the eye movement measures, qualitative differences can be seen in the graphs of the means (Figures 1-6). This implies that there is the potential that eye movement measures can be used to help further understand the usability problems indicated by performance measures. That is, eye movement measures appear to be sensitive to similar influences as the more conventional performance/usability measures.

An important issue that arises from both this study and those of Goldberg and Kotval is that the usability of the interfaces tested was only relative to the other interfaces. There is no absolute benchmark of what eye movement patterns a good or a bad interface produces. Also, individual differences in participants' performance make it difficult to say that a certain pattern of eye movements should occur for an interface to be of a particular standard. For example, doing Task 1 on Page A, one participant had a mean fixation duration of just 297ms, whilst another's was 738ms; one took only 17s to complete that task on that page, another took 42s.

Several issues have arisen from this exploratory study that would benefit from further research. First, it is necessary to establish a means of identifying benchmarks for eye movement patterns on a single interface design, rather than just for comparing a selection of interfaces. Second, it is important to investigate the types of processing and search strategies that are employed by users on interfaces and the eye movement patterns associated with them. It can then be considered how eye movement measures can be adapted to detect them more accurately. Third, it is probable that the number of possible correct targets on the interface affects the usability of the page. The Response Scores in the present experiment suggest that they may have some influence, but that the visibility of those possible targets is also important. Finally, it is possible that using a greater variety of tasks would reveal a greater sensitivity of the eye movements to various factors that affect the usability of an interface.

References

B R Buckingham.
New Data on the Typography of Textbooks
Yearbook of National Society for the Study of Education 30
Part II The Textook in American Education
93-125
1931

T Baccino
T Colombi
L'Analyse des Mouvements des Yeux sur le Web.

Les Interactions Homme-Système: Perspectives et Recherches Psycho-ergonomiques
A VomHofe
Hermès
127-148
2001

E C Crowe
N H Narayanan
Comparing Interfaces Based on What Users Watch and Do
Eye Tracking Research & Applications Symposium 2000
29-36
2000

J H Goldberg
X P Kotval
Eye Movement-based Evaluation of the Computer Interface
Advances in Occupational Ergonomics and Safety
S K Kumar
IOS Press
529-532
1998

J H Goldberg
X P Kotval
Computer Interface Evaluation Using Eye Movements: Methods and Constructs
International Journal of Industrial Ergonomics
24
6
631-645
1999

J H Goldberg
X P Kotval
Eye Movement-Derived Measures of Interface Usability
In review
1999
Retrieved March 2001 from the World Wide Web:
http://www.ie.psu.edu/people/faculty/goldberg.htm

R L Gregory
Eye and Brain: The Psychology of Seeing 4th edition
Weidenfield and Nicholson
1990

J M Henderson
A Hollingworth
High-level Scene Perception
Annual Review of Psychology
50
243-271
1999

R J K Jacob
Eye-Tracking in Advanced Interface Design
Virtual Environments and Advanced Interface Design
W Barfield
T A Furness
Oxford University Press
258-288
1995

X P Kotval
J H Goldberg
Eye Movements and Interface Component Groupings: An Evaluation Method
Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society
HFES
17-23
1998

J Nielsen
Usability Inspection Methods
Heuristic Evaluation
J Nielsen
R L Mack
John Wiley and Sons
25-62
1994

S E Palmer
Vision Science: Photons to Phenomenology
Massachusetts Institute of Technology
1999

J Preece
H Sharp
D Benyon
S Holland
T Carey
Human–Computer Interaction
Addison-Wesley Publishing Company
1994

K Rayner
Eye Movements and Cognitive Processes in Reading, Visual Search, and Scene Perception
Eye Movement Research: Mechanisms, Processes and Applications
J M Findlay
R Walker
R W Kentridge
Elsevier Science B. V.
3-22
1995

SensoriMotoric Instruments
iView version 3.01 Manual & Software Reference Document version 3.01
SMI
1999

C Wharton
J Rieman
C Lewis
P Polson
The Cognitive Walkthrough Method: A Practitioner's Guide.
Usability Inspection Methods
J Nielsen
R L Mack
John Wiley and Sons
105-140
1994