Eye tracking in usability evaluation: A practitioner's guide

Book · January 2003

CITATIONS
48

READS
4,210

2 authors:

Joseph H. Goldberg
self employed
76 PUBLICATIONS **2,296** CITATIONS
SEE PROFILE

Anna M. Wichansky
Oracle Corporation
26 PUBLICATIONS **454** CITATIONS
SEE PROFILE

Eye Tracking in Usability Evaluation: A Practitioner's Guide

Joseph H. Goldberg and Anna M. Wichansky

Oracle Corporation
Advanced User Interfaces
500 Oracle Parkway
MS 2op2
Redwood Shores, CA 94065

March 2002

[To appear in: Hyönä, J., Radach, R. and Deubel, H. (Eds.), The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements, Elsevier Science, Oxford, 2002.]

ABSTRACT

This chapter provides a practical guide for either the software usability specialist who is considering the benefits of eye tracking, or for the eye tracking specialist who is considering software usability evaluation as an application. The basics of industrial software usability evaluation are summarized, followed by a presentation of prior usability results and recommendations that have been derived from eye tracking. A detailed discussion of the pitfalls and methodology of eye tracking is then provided, focusing on practical issues. Finally, a call for research is provided, to help stimulate the growth of eye tracking for usability evaluation.

1. USABILITY EVALUATION

Usability evaluation can be defined as any of several techniques where users interact with a product, system, or service, and some behavioral data are collected (Wichansky, 2000). It is conducted, within many industries, before, during and after development and sales of products and services to customers. This is a common practice in computer hardware, software, medical devices, consumer products, entertainment, and on-line services such as electronic commerce or sales support.

What techniques are used for usability evaluation?

There is wide variation among academic and industry implementations of usability evaluation techniques. Some techniques are very informal, providing little more than nominal or qualitative data, and others are methodologically very rigorous, involving hundreds of participants and highly controlled administration protocols. The most frequently used techniques are various forms of usability testing, which have in common several key characteristics:

- Users are selected from target market groups or customer organizations.
- Users interact systematically with the product or service.
- They use the product under controlled conditions.
- They perform a task to achieve a goal.
- There is an applied scenario.
- Quantitative behavioral data are collected.

Other techniques which partially fulfill the above requirements are known as walkthroughs, feedback sessions, and various other terms. Often they lack the participation of "real users", they are not controlled, they may not provide a task, and there may be no systematic collection of quantitative data. The results of these *formative usability tests* may not answer the question, "Is this product usable?," but can be of value in making design decisions about a product under development.

The Industry Usability Reporting Project, an industry consortium led by the National Institute of Standards & Technology (NIST), has recently developed a standard for reporting usability testing data (Scholtz et al, in press). This standard is particularly useful when customers need to make procurement decisions about a supplier's software. This Common Industry Format (CIF) is available as ANSI/NCITS 354-2001. The requirements necessary for CIF compliance were based upon the human factors literature, existing ISO standards on usability (e.g. ISO 9241-11 and ISO 13407), and industry best practice. They include requirements for goal-oriented tasks, minimum of 8 participants per user type, and quantitative metrics representative of efficiency, effectiveness and satisfaction. *Summative testing* techniques, such as those proscribed by NIST, provide a better answer than the formative techniques to the question of whether a product is usable.

What software is tested?

In large software companies, there can be so much software developed and manufactured, that insufficient resources are available to test all of it for usability. Typically, new products, those undergoing drastic changes, and those for which customers have provided negative feedback are prioritized higher for usability evaluation. Successive versions of the same product may contain new features to be tested in the context of the original product in development. Where the user interfaces are the same, there is typically less emphasis on testing. In the realm of Web software, it is difficult to conduct controlled tests because websites change frequently, not only in content, but in basic architecture. Therefore, any test is really a snapshot in time of the product's usability.

When is software tested for usability?

Software should be tested early and often throughout the product development cycle. In this way, problems with the product's conceptual model can be identified early, before too much coding has been completed and it is too late to change the product. This is done by developing prototypes that represent the product's functionality and interface design early in the product development cycle. These prototypes may range in fidelity from low

(paper drawings) to medium (partially functional screen representations with little or no behavior) to high (interactive screen renditions of the product's front-end, with partial or simulated functionality). As the product is developed, alpha code, representing feature functionality, and beta code, representing complete product functionality, can be tested. At or near product release, summative testing as per the NIST CIF can be conducted. At every successive stage, users can be presented with more challenging tasks, and more definitive feedback and recommendations can be provided to the developers for changes to make the product more usable.

Who conducts usability tests?

Usability practitioners come from a wide variety of backgrounds. Typically they have some expertise or education in experimental psychology and statistics. They often have degrees in industrial engineering, psychology, ergonomics, or human factors engineering, and a familiarity with topics such as laboratory and field data collection, and the overall scientific method. Many companies employ user interface designers and usability engineers with some or all of the above background. In other organizations, individuals with little or no formal training conduct usability tests. These team members may have developed a strong interest in usability through experiences in product management, technical writing, customer support, or other related roles.

What metrics are associated with usability testing?

Most industry tests include some subset of behavioral metrics that can be identified with efficiency, effectiveness and satisfaction. Complete descriptions of these aspects of usability and related metrics are available in Macleod et al (1997). *Efficiency* is similar to the concept of productivity, as a function of work accomplished per unit time. Common measures are time on task (often weighted for errors or assists from the test administrator) and mean completion rate : mean task time, known as the efficiency ratio. *Effectiveness* is how well the user is able to perform the task, in terms of the extent to which the goal is accomplished. A common measure is percent task completion, again weighted for errors, assists, or references to support hotlines or documentation. *Satisfaction* is determined from subjective measures that are administered at task completion, or at the end of the usability test. These include rating scales addressing concepts such as usefulness, usability, comprehensibility, and aesthetics. Psychometrically developed batteries of subjective questions are often used, such as the Software Usability Measurement Inventory (Kirakowski, 1996) and the Questionnaire on User Interface Satisfaction (Chin et al, 1988).

What is a typical test protocol?

Typically test participants are brought individually into a usability testing lab, which consists of a control room with a one-way mirror, and a user room that is often designed as a simulated usage environment appropriate to the product (Figure 1). After receiving material on the rights of human test participants (Sales and Folkman, 2000) and signing consent forms, the user is presented with a set of tasks, which may be read aloud or

silently. The tester sits behind the one-way mirror, unseen by the user, and may call out instructions to begin or end tasks as appropriate. A think-aloud technique, in which the user describes out loud his mental process as he performs the steps of the task, is often used (Dumas, in press). Think-aloud protocols are particularly useful in the formative stages of product development, when there is a lot of latitude for making changes. In summative testing, the user may perform the task silently, and merely indicate when he is done. All of this is usually captured on videotape, for possible later review. The tester typically logs various aspects of the user's behavior, including task start and end times, times to reach critical points in the task, errors, comments, and requests for assistance. Questionnaires are usually administered at the end of the test, and the user may be interviewed about his experience.

Figure 1. Tester instructs participant in a usability lab at Oracle.

How is the data from a usability test exploited?

The data from formative tests are used to determine where the product's usability is satisfactory and where it needs improvement. If certain tasks are performed particularly slowly, generate many errors and requests for assistance, or produce negative comments, the user interfaces supporting those features and functions are analyzed to determine the cause of problems. Then they are corrected, usually in conjunction with the user interface designers and developers.

The data from summative tests are used to assess overall usability with reference to goals that have been established by the development team. For example, a particular task may need to be performed by the average user in a set period of time with no errors, or with no outside assistance. The data also make it possible to compare the current product to previous versions, or to other products.

What aspects of usability cannot be addressed by routine usability testing?

Current usability techniques are derived from industrial engineering, which concerns itself with the measurement of work in the workplace, and the behaviorist school of psychology, which emphasizes operational definitions of user performance rather than what may be inferred about the cognitive process. Thus, most observable, behavioral aspects of work and task performance (e.g. screen navigation, menu selection) can be captured. Cognitive processes are more difficult to infer.

User intent is difficult to assess using current techniques. If users spend too long looking at a specific window or application page, for example, traditional usability analysis does not provide sufficient information for whether a specific screen element was found, or whether its meaning was unclear (Karn et al, 2000). Micro-level behaviors, such as the focus of attention during a task, distractions, or the visibility of an icon, usually have little awareness to an individual, and thus are not reported in the think-aloud protocol. Reading, mental computations, problem solving, and thinking about the content of an application are also difficult to quantify using this protocol.

There are techniques aimed at assessing mental models and user intentions, such as cognitive walkthroughs (e.g., Polson et al, 1992) and protocol analyses (e.g., Card et al, 2001). These dwell upon thought processes, but may be very time and labor-intensive and are often practical only in research settings.

Usability testing specialists would like better tools to understand when users are reading versus searching on a display, and to determine how learning a screen's layout impacts its usability. Current evaluation approaches can capture keystrokes and cursor movements, allowing inferences to be made about complex cognitive processes such as reading, but cursor motion does not necessarily track where the user is looking (e.g., Byrne et al, 1999). Thus, better tools are required to assess temporal and sequential order of viewing visual screen targets, and to assess when users are reading as opposed to searching on a display. Eye tracking can provide at least some of this data.

What types of design recommendations have been made from eye tracking results?

<u>User interface architecture and screen design</u>

The outcome of using eye tracking as a secondary technique in usability studies should be improvements in the breadth and specificity of design recommendations. Most studies have collected data relevant to navigation, search, and other interaction with on-line applications. These findings influence overall user interface architecture of software applications, design and content of screens and menus, location and type of visual elements, and choice of user interface visualization style.

Recommendations for icon size for normal and vision-limited users were provided by Jacko et al (2000), by requiring users to search for icons of varying sizes. Based upon

fixation and scanpath time durations, they recommended minimum icon sizes that varied by background brightness on displays.

Menus contain both reading and visual search subtasks, and provide a well-learned interaction model for eye tracking purposes. Crosby and Peterson (1991) investigated scanning style as users searched multi-column lists for specific target items. Four styles were uncovered from eye tracking: Comparative (scanning between columns and/or rows to compare specific items); Down and Up (starting from the top of a column and continuing from its bottom to the bottom of its neighboring column, then back up to its top); Scan from Top (always scanning downward from the top of each column); and Exhaustive (scanning all areas and columns, even after finding the target item). These strategies were related to an independently measured index of cognitive style.

Because menu search can vary from highly directed to an informal browsing activity, Aaltonen et al (1998) manipulated the directness of the search target through instructions. From eye tracking results, either top-down scanning of each item or a combination of top-down and bottom-up scanning strategies were evident. Analysis was not conducted, however, to predict the stimuli that drive these strategies. In similar work, Byrne et al (1999) had users search for target characters in pulldown menus while menu length, menu target location, and target type were manipulated. Menus of six items produced a similar number of fixations, regardless of item location within the menu. Initial fixations were made to one of the first three menu items, with subsequent search occurring in serial order.

Pirolli et al (2001) used eye tracking as an aid to interpret users' behaviors in navigating hyperbolic and other network browsers. Based upon the number and duration of fixations, they recommended a hyperbolic browser in directed search tasks, as visual link traversal rate was nearly twice as fast as a traditional tree browser.

Redline and Lankford (2001) provided a very detailed analysis of users' behaviors while using an on-line survey. They studied scanpaths from eye tracking data in order to categorize users' omission or commission errors. Example behaviors included failure to observe branching instructions, premature branching from one question to another, or non-systematic fixations leading to a non-answer. Overall, the authors noted that the order in which information was read influenced whether the survey was understood as intended.

Goldberg et al (in press) evaluated a portal-style website with eye tracking techniques, allowing users to freely navigate within and among web pages. They recommended that portlets requiring the most visibility be placed in the left column of the portal page; 'Customize' and 'More' links found in many portlet headers may benefit from being added to the bottom of the portlet body. In a much earlier evaluation of a Prodigy web browser, Benel et al (1991) found that users spent much more time than designers anticipated looking at special display regions.

Reading

Eye movements during reading are different from those exhibited during navigation and search of an interface. The eyes follow a typical scan path across the text, in the appropriate direction depending upon the language of the text (e.g. left to right, top to bottom for Western cultures). This is a highly learned behavior. When there are difficulties with legibility or comprehensibility of the text, the eyes may dwell on specific words or samples of text, or even reverse direction to reread text, slowing down reading progress and disturbing the normal pattern.

Reading research has presented a rich eye tracking environment for many years. As a task, reading is much more constrained and context-sensitive than spatial/graphical visual search. Reading pattern differences are clear between novices and experts, and complexity of written material can be discerned from eye tracking data. Rayner (1978) provided several facts about eye movements in reading. Reading consists of a series of short (saccadic) eye movements, each spanning about 8 characters (with a range of 2-18 characters) for average readers. Regressions, marked by right-to-left eye movements, occur about 10-20% of the time for skilled readers, and more often in novice readers. A very distinctive return sweep, which is distinguishable from a regression, is made as one reads from one line to the next. Textual information in the periphery of the eye is also considered to be important for gaining context and preparing the next several saccades.

Eye tracking studies for reading tasks have provided a great deal of information concerning the lag between text perception and cognitive processing. For the usability specialist, eye tracking can provide information about when a participant is searching for visual targets versus reading text. Given sufficient textual material on a display, eye tracking can also provide an indication of the participant's difficulty in reading that information. However, there are strong individual differences in reading, so studies should preferably present within-participant manipulations.

Studies involving reading of on-line documents coupled with eye-tracking methods are indeed quite relevant to usability evaluation. Zellweger et al (2001) designed a variety of "fluid document" strategies to enable supplementary text (similar to footnotes) to be presented in on-line documents. They evaluated four hypertext designs using eye-tracking: fluid interline (lines part, hypertext appears in between lines); fluid margin (hypertext in the margin); fluid overlay (hypertext over text); and pop-up (hypertext in a separate box). From eye tracking data, participants often did not look at footnotes located at the bottom of the document. Application of these "glosses" did not cause wildly shifting point of regard, as anticipated by the authors. Eye tracking data was confirmed by subjective results and by amount of time glosses were opened: shorter for glosses farther from the anchor text and longer for glosses local to the anchor text.

In 1998, researchers at Stanford University Center for the Study of Language and Information began to collaborate with the Poynter Institute (2000), dedicated to journalism research, in a study of on-line newspaper reading behaviors. Eye-tracking measures were used in addition to videotaped observation and survey measures with 67

participants in two cities. Among the results, readers initially looked at text briefs and captions, not photographs or graphics, upon navigation to a news-oriented website. Eye tracking revealed that text is sought out and either skimmed or read. Banner ads were read 45% of the time and fixated an average of 1 sec, long enough to be clearly perceived. Readers scanned 75% of the length of most articles. The authors found the eye-tracking data to be an important addition to surveys and recall data.

Cognitive Workload

Although still under active investigation, eye tracking can possibly provide significant information about an observer's perceived cognitive workload. Several studies have presented tasks in which participants must actively control and make time-limited decisions about a system. Various eye tracking-derived metrics are generated, and possibly correlated with subjective scales of perceived workload. As cognitive workload increases, the saccadic extent, or breadth of coverage on a display, decreases (e.g., May et al, 1990), consistent with a narrower attentional focus. The blink rate of the eye also decreases (Brookings et al, 1996), emphasizing the need to gather as much visual information as possible while under high workload. Although influenced by many factors, some have found that the pupil dilates in response to high workload tasks (e.g., Hoeks and Levelt, 1993). In response to difficult cognitive processing requirements, fixation durations may also increase (Goldberg and Kotval, 1999). The usability specialist is advised to use caution before interpreting eye tracking-derived measures as indicators of workload. Large individual differences and the presence of many contaminating factors could lead to false conclusions. Within-participant manipulations are advised here, noting relative differences between conditions.

How do eye tracking results correlate with other measures of usability?

In studies where this has been reported, eye tracking data typically support other measures of usability. For example, Kotval and Goldberg (1998) investigated the sensitivity of eye tracking parameters to more subjective usability evaluations by interface designers. They developed several versions of a simulated drawing package, by manipulating items' grouping and labeling on the application's tool bars. Screen snapshots were then sent to a broad set of interface designers in the software industry for subjective rating and ranking. Participants in the eye tracking laboratory completed several tasks using the software, and metrics from eye tracking were then correlated with the designers' subjective ratings. Several parameters from eye tracking were indeed related to the usability ratings. Cowen (2001) extended Kotval and Goldberg's (1998) work to 'real' web pages, using eye tracking as an aid to evaluating the usability of two tasks on each of four web pages. Differences in rated usability and completion time on the web pages did indeed covary with eye tracking measures such as fixation duration, number of fixations, and spatial density of fixations.

Goldberg (2000) discussed which traditional usability criteria can be addressed via eye tracking methods. Understanding this can more effectively determine which aspects of usability assessment can be aided by eye tracking. Eye tracking should generally be an

excellent indicator of visual clarity of an interface. Criteria that could be ascertained somewhat from eye tracking include cognitive resources and flexibility of use. Criteria that might be related to eye tracking, but to a limited extent, include feedback and error handling. Finally, criteria that would be difficult to ascertain from eye tracking include interface compatibility, and locus of control. The above should provide general guidelines, but should not be blindly used; specific interface features and eye tracking details will certainly modify these generalities.

2. EYE TRACKING BASICS

The usability specialist who seeks to add eye tracking to usability evaluations will require some basic knowledge about eye movements and eye trackers. The following information should help, but this information is by no means exhaustive. For example, types of eye trackers other than those mentioned do exist, but are not common in the domain of user interface evaluation.

Which types of eye movements are measured in usability evaluations?

Eye movements exist in humans to either maintain or shift an observer's gaze between visual areas, while keeping the projected visual image on a specialized, high acuity area of the retina. There are several types of eye movements, based upon their specific function and qualities. Young and Sheena (1975) and Carpenter (1988) provided introductory reviews.

Saccades are commonly observed when watching an observer's eyes while conducting search tasks. These jerky movements occur in both eyes at once, range from about 2-10 degrees of visual angle, and are completed in about 25-100 msec. They have rotational velocities of 500-900 degrees/second, and so have very high acceleration (Carpenter, 1988). Saccades are typically exhibited with a latency of about 250 msec following the onset of a visual target. Because of their rapid velocity, there is a suppression of most vision during a saccade to prevent blurring of the perceived visual scene.

Each saccade is followed by a fixation, where the eye has a 250-500 msec dwell to process visual information. These saccade-fixation sequences form scanpaths, providing a rich set of data for tracking visual attention on a display (Noton and Stark, 1970). Applied eye tracking methods for usability evaluation generally capture saccadic eye movements and fixations while observers search displays.

Smooth Pursuit eye movements are much slower (1-30 degrees/second) movements used to track slowly and regularly moving visual targets. These cannot be voluntarily generated, and may be independently superimposed upon saccades to stabilize moving targets on the retina.

Compensatory eye movements stabilize the image on the retina when the head or trunk is actively or passively moved. These slow eye movements occur in the opposite rotational

direction from the head or trunk movement. They do not usually occur in usability evaluation studies, because of little head or body movement.

Vergence eye movements are rotational movements of the two eyes in opposite directions, in response to focusing error from visual targets moving closer or further away. The direction of rotation depends on whether a visual target is moving closer or further away from the eyes. These slow, 10 degree/second eye movements are necessary to fuse the two images from the eyes. They are usually not observed in computer-oriented studies, as the viewing plane does not typically change.

Miniature eye movements refer to those eye movements that are under 1 degree of visual angle, and include small corrective motions, as well as tremor. These are usually not of interest in usability evaluations, as eye trackers typically have accuracy error of this magnitude.

Finally, *nystagmus* eye movements refer to a unique sawtooth-pattern of eye movements caused by a rapid back-and-forth movement of the eyes. These movements occur for several reasons, such as fatigue, hot or cold water in the ears, rapid rotation of head, or special optokinetic stimuli. They are not typically observed in usability evaluations.

What data is gathered by an eye tracking system?

Eye tracking methods attempt to capture a user's focus of visual attention on a display or scene through special hardware/software. Hardware may be attached to the user ('head mounted' systems), or may be as non-intrusive as a small camera near a display ('remote' systems). The eye tracking system gathers x/y location and pupil size information at typical rates from 30 Hz – 250 Hz, with 60/50 Hz typical for usability evaluation needs. Slower sampling rates do not provide sufficient resolution of visual attention, especially for scrolling or other tasks involving moving visual targets. Faster sampling rates can create a massive data reduction problem for typical usability evaluations. The x/y sample locations are then further processed into fixations, which may be assigned to experimenter-defined areas-of-interest on the viewed display or scene. Fixations and saccades can be combined into scanpaths, providing further information about one's cognitive processing during a task (Just and Carpenter, 1976b). Scanpaths can be quantitatively or subjectively analyzed to provide information about the extent or complexity of visual search on a display (Goldberg and Kotval, 1999; 1998).

What do eye trackers track?

Eye tracking systems can use many different properties of the eye to infer an observer's gaze angle. Several methodologies utilize reflection and refraction of a small light glint, as it is reflected from and refracted through the eye. Young and Sheena (1975) included other characteristics that may be used:

- light reflection from special contact lenses
- video images of the pupil and/or iris
- brightness differences between the iris and sclera

- microscopic images of the retina
- a small electrical dipole between the lens and the retina
- electrical impedence changes between electrodes placed near the eyes
- the bulge produced by the cornea

There are advantages and disadvantages for each of the above; some are more accurate and/or reliable, but may be more intrusive. Some can result in relatively inexpensive eye trackers, but may have poor accuracy. Overall, corneal reflection, video-based systems now dominate the applied eye tracking market. These combine video imaging technology with small camera lenses and an infrared light source.

How does an infrared, corneal reflection system work?

Infrared-type, corneal reflection eye tracking systems, in common use for applied research, rely upon the location of observers' pupils, relative to a small reflected light glint on the surface of the cornea (Young and Sheena, 1975, Mulligan, 1997). A camera lens (the 'eye' camera) is focused upon the observer's eye; a second lens (the 'scene' camera) may also optionally be pointed towards the current visual display or scene being viewed. A scan converter is frequently used in place of the scene camera when tasks are conducted on a computer. The pupil is located by the tracking system (and possibly modeled as an ellipse) from its relative contrast with the surrounding iris. 'Bright pupil' systems illuminate the pupil with infrared light, whereas 'dark pupil' systems do not illuminate the pupil; some systems may be switched between these modes, to find the most robust pupil imaging for a testing environment. Within a usability laboratory, diffuse lighting will generally provide fewer problems that pinpoint lighting sources that might cause additional light glints on the cornea. The ability to switch between bright and dark pupil methods can also be helpful in usability testing situations.

Following calibration to the display or scene of interest, the eye tracking system returns x/y coordinates of gaze samples, an indication of whether the eye was tracked, and pupil size. These data may be streamed and used in real-time, or may be saved to a file for later analysis. Note that modern eye tracking systems come with software/hardware that automatically filters out blinks and other anomalies. Embedded algorithms within commercial eye tracking software can detect and filter out blinks, by searching for unique light-dark-light signatures of reflected light that are associated with blinking. Special blink detection hardware is available for blink rate and fatigue research (e.g., Stern et al, 1994).

What are the pros and cons of head-mounted and remote systems?

Eye tracking systems may be mounted to the head of an observer, or may be remote from the observer. Head mounted systems (e.g., Figure 2) are most useful for tasks in which a great deal of free head/trunk motion is expected. Eye tracking during sports, walking, or driving are common examples. While head mounted systems can continue to track the eyes even with significant head movement, they are somewhat instrusive, costly, and

delicate. They typically obstruct a small portion of the observer's visual field, and the observer cannot easily forget that a system is recording his eye movements.

Remote systems, on the other hand, consist of a camera lens mounted under or beside the computer display being used by the participant in a usability evaluation (Figure 3). An infrared source is typically located in-line with the camera lens. The presence of the camera is quickly forgotten by the participant, making this a very non-intrusive procedure. In fact, special systems are available that can even hide the camera lens entirely from view. These systems can be a bit less expensive and are less delicate than the head mounted systems, but require the participant to maintain relatively stable head position. Chin rests could be required if head motion is not easily controlled. Eye cameras that have an autofocus capability allow some forward-backward movement of the participant's head, an additional advantage for usability testing.

Modern, remote eye tracking systems allow head motion within approximately one cubic foot, by coupling a motor/servo driven platform and magnetic head tracking with the eye camera, enabling robust location of the eye as the participant sits naturally in front of a computer. The addition of these head tracking technologies can make the cost of a remote system similar to that of a head mounted system. Costs over the past few years have ranged from about $20,000 - $50,000 for a complete eye tracking system; however, new systems in the sub-$20,000 range are now starting to appear.

Whether head-mounted or remote, the eye tracker generates data that is fed to the port of a host computer, possibly through a control box. This computer is usually dedicated to the eye tracker, containing a large amount of disk space for data storage. Many investigators use a two-computer setup, where software under evaluation is run on an independent, application computer. The experimenter can control the eye tracker data collection through the host computer. Optionally, the host computer can also be programmed to send signals to the eye tracker to start or stop tracking. These signals could also be generated from the

Figure 2. Head-mounted eye tracking system, including both scene (lower) and eye (upper) cameras.

Figure 3. Remote eye tracking system, with camera optics below and to the right of the computer display.

application computer, in response to user-generated events such as mouse movements or button presses.

What steps are required for analyzing eye tracking data?

Analysis of eye tracking data typically requires several steps, moving from raw data to fixations to computation of specific metrics. Reduction of eye tracking data into meaningful information can be a tedious process. Summaries of this process have discussed in detail the steps that are required (e.g., Goldberg et al, 1999; Jacob, 1995).

Raw data samples are usually aggregated off-line by the investigator into meaningful behavioral units of fixations and saccades. Although the temporal and spatial requirements that define these vary among investigators, the intent is to quantitatively describe the behavioral tendency of an observer's attentional flow on an interface. The fixations are written into a new file, listing start/end times, and locations. Additionally, a code may be included, corresponding to the quality of the original signal (i.e., tracking status) received by the eye tracker.

Because fixations are made at a rate of approximately 3 Hz, the datafile length is reduced by a factor of about 20. While reduction to fixations is typically automated by software, the investigator is strongly encouraged to review the created fixations against images of viewed displays to ensure that the fixations are valid. Further processing of the fixation file can identify which defined screen objects were captured by each fixation, and thus compute the instantaneous and cumulative dwell times by screen objects. Sequences and transitions among these objects can then be computed. Goldberg and Kotval (1999; 1998) provided a detailed discussion and classification scheme for these higher level metrics. Commercially-available software has also been developed to aid the data reduction process (e.g., Lankford, 2000).

3. METHODOLOGICAL ISSUES IN EYE TRACKING

There are both advantages and disadvantages to eye tracking methods for usability evaluation. The ability to record one's micro-flow of visual attention in a non-intrusive way is certainly an advantage, but this can create a huge, tedious data reduction problem. Individual eye movements are quite randomly distributed, often requiring inferential statistics for discovering scanning trends. Strong individual differences are evident, ranging from individuals who scan broadly to those who barely make observable eye movements. Hardware must be frequently calibrated, and subtle differences in eye color or eye kinematics can cause an eye tracking failure. The usability specialist must certainly weigh these issues before deciding to implement an eye tracking methodology within a usability study. The present section considers many of these methodological issues that should be understood by the usability specialist. While ignoring one or more of these may not invalidate a study, the tester might want to alter an experimental procedure to better accommodate some of these pitfalls.

If eye tracking is desired, but it is not possible to purchase or rent a system, consulting firms exist which provide evaluation services. These companies can provide on-site eye tracking services, reduce data, and provide a report. Before hiring these services, however, it would be wise to understand the scope of evaluation that will be desired, and to provide a list of questions/hypotheses that the eye tracking service needs to answer.

Does the eye's current location indicate what one is currently thinking about?

Eye tracking methods generally rely upon the *eye-mind hypothesis*, which states that, when looking at a visual display and completing a task, the location of one's gazepoint corresponds to the thought that is 'on top of the stack' of mental operations (Just and Carpenter, 1976a). Visual attention may, however, lead or lag the current gazepoint, under certain conditions (Rayner, 1978). In reading text, for example, rapid eye movements followed by short fixations cause one's semantic processing of text to lag behind perceptual stimulus input. Reading difficult text passages very quickly leads to longer fixation durations and re-reading of passages. Certain task factors can also decouple the eyes from the mind, reducing the sensitivity of eye movements as indicators of mental processing. Examples of these factors include spatial and temporal uncertainty about where important information is located, low visual target salience, task interruptions, and a high peripheral visual load (Potter, 1983).

Although there are no guarantees that the eye-mind hypothesis will be valid, several steps can be taken to support this hypothesis insofar as possible (Just and Carpenter, 1976a):

- *Tasks should be designed to require the encoding and processing of visual information to achieve well-specified task goals.* Tasks in which observers are only requested to 'look at' a page or scene don't provide these necessary task goals. Rather, goal-directed instructions to search for (and possibly read and process) a particular item within a specified time limit should be specified.
- *Extraneous peripheral information should be controlled within tasks.* Screens could be masked off, or highly distracting blinking, moving, or colorful peripheral items could be eliminated.
- *Scanning uncertainty should be minimized within tasks.* Observers should have a general notion of the location of visual targets, and should not be surprised by the location of target items.
- *The 'behavioral unit' of experimental tasks should be large enough to be open to conscious introspection.* Eye fixations generally correlate well with verbal protocols when decision makers are solving problems or choosing from among several alternatives. For example, selection of a menu item represents a larger behavioral unit than reading specific words within each menu item.
- *Cross-modality tasks may be candidates for eye tracking methodologies.* When listening to spoken linguistic passages, for example, the eyes tend to fixate upon spatial or pictorial referents of the words in the text.

How does the screen cursor influence one's visual attention?

Because a moving cursor can provide a highly salient and attention-drawing visual target on a display, it is important to understand the impact of cursor movements on scanpaths. In their study of menu item selection, Byrne et al (1999) noted that the eye initially fixates a visual target, and is then trailed by cursor movement. Smith et al (2000) had participants use a mouse, touchpad, or pointing stick to perform either a reciprocal or random pointing task on a display. Most participants' visual attention either led or lagged the cursor, but some exhibited frequent target-cursor-target switching. The eyes tended to lead the cursor for known target locations, but lead vs. lag strategies were difficult to predict, based upon task factors.

Does task contrast and luminance influence eye tracking?

In the case of infrared eye tracking systems, very small pupil sizes can make it difficult for the eye tracking system to model the pupil center, causing poor eye tracking. Also, very small pupils are more easily hidden by the lower eyelid, especially if the eye tracking camera is poorly positioned. The pupil contracts and expands in response to the overall luminance and contrast of an observed display. Tasks presenting screens with large bright areas (e.g., lots of white space) cause the pupil to be smaller than those that contain screens with darker backgrounds. Tasks that contain animated images, scrolling, extreme foreground:background contrast, or extreme variance in luminance cause the pupil size to frequently vary. There are several recommendations to ensure the best possible eye tracking conditions for a task:

- *Minimize the use of pinpoint task lighting.* This can cause secondary glints on the observer's corneal surface, confusing the eyetracking system. There should also be no bright, reflective surfaces in the room, which could cause inadvertent light glints.
- *Avoid bright lighting and/or bright task screens.* To ensure an observer's pupils are sufficiently large for the system to locate and model, the testing room should ideally be dimly lit and screens should be uniform in background luminance. Because usability testing labs often use bright lighting to accommodate video cameras, some compromise may be necessary. Both room lights and the infrared source on the eye tracker should be controlled by variable rheostats. Note that, if luminance is highly variable on test screens, increasing room luminance may somewhat negate the screen influences to pupil size changes, enabling a steadier pupil size.
- *Ensure proper geometry between the eye tracking camera and the observer's eye.* Whether remote or head-mounted, the eye tracking camera is generally setup to view the eye from a position that is lower than the observer's horizontal horizon. If mounted too low, the lower eyelid can obstruct proper pupil imaging, especially for small pupils.

Are large differences between individuals expected when conducting eye tracking?

Informal estimates suggest that as high as 20% of recruited participants will have problems with loss of tracking or calibration on an eye tracking system. For instance,

Crowe and Nayayanan (2000) found that participants' eyes were not tracked 18% of the time for a keyboard/mouse interface; this increased to 35% of the time for a speech interface, when more head movement was noted. Excessive head movement during and between calibrations is quite problematic for remote eye tracking systems. Those with extremely dark iris colors may not have sufficient contrast between pupil and iris for the system to find the pupil's center. Some individuals have lower eyelids that cover substantial portions of the pupil and/or iris.

Glasses and contact lenses (soft or hard lenses) are frequently problematic to effective eye tracking. Lenses refract both the incident source and the reflected infrared glint from the cornea, possibly causing high-order non-linearities in calibration equations, especially for extreme right or left eye rotations. Both types of lenses may also cause secondary light glints on the front or back sides of the lenses, making it difficult for the system to determine which light glint is associated with the corneal surface. Contacts have the possible additional problem of slippage as the eye rotates. The following suggestions may alleviate some of these tracking problems:

- *Minimize head motion.* Ensure that observers' heads remain as stationary as reasonably possible during calibration procedures, and if required, during testing. Chin rests or bite bars may be necessary in extreme cases.
- *When confronted with an individual with extremely dark irises, try both bright and dark pupil tracking methods, if possible.*
- *Pay careful attention to camera setup.* When setting up the eye tracking camera relative to the observer's position, be careful that the lower eyelid does not overly obstruct the view of the iris and pupil.
- *Consider limiting glasses and contact lens wearers during recruitment*. Participants with weaker prescriptions, for example, may calibrate more successfully than those with stronger prescriptions.
- *Over-recruit for participants*. Recruit up to 20% more participants than required in case some fail to calibrate to the eye tracker.

Which eye is tracked?

Most eye tracking systems for behavioral analysis only track a single eye. Though binocular tracking systems are available, these are only required when vergence (or any other non-conjugate) movements must be measured, as in frequent, rapid changes in distance between observed targets and the eyes. Tracking of a single eye is usually adequate for most tasks, as the eyes are yoked when tracking objects moving across one's visual field. Because the light source is typically infrared, the participant has little awareness of the eye tracking apparatus, and can see stimuli clearly from both eyes.

Most head-mounted and remote eye tracking systems allow tracking of either eye. Generally, the investigator determines the dominant eye of the participant, and uses that throughout testing. Approximately 75% of the population has eye dominance on the same side as their hand dominance; the rest are either opposite-dominant, or exhibit no eye dominance. Eye dominance may be determined by a simple test:

- The participant is asked to point to a far object with an outstretched arm, using both eyes.
- While still pointing, the participant is asked to close one eye at a time.
- The eye that sees the finger pointing directly at the target is dominant.

How is eye tracker calibration and maintenance conducted?

Calibration procedures are a very necessary and critical part of eye tracking studies, in order to relate an observer's gaze angle, or point-of-regard, to locations in the environment (e.g., computer display). Poor calibration can invalidate an entire eye tracking study, because there will be a mismatch between the participant's point-of-regard, and the corresponding location on a display. Hardware drift and head movement are two big problems that can invalidate a set of data.

Calibration should be conducted multiple times per user session, as well as between users. Commercial eye tracking systems generally come with calibration software that may be run independently or inserted into one's experimental test code. Building a calibration routine within experimental stimulus presentation code is a good idea, especially if the calibration can easily be conducted on demand (e.g., between trials or blocks of trials). Recalibration is recommended every few minutes, such as between trial blocks.

The calibration procedure generally requires an observer to successively fixate a series of known locations on a display, or within the environment. On a display, these locations are usually near the corners, and at the display's center in a 5-point calibration. A 9 or 13-point calibration locates these at the vertices of successively larger squares, starting from the display's center. Note that random location calibrations are also possible, but may require more locations to obtain a reliable calibration model. It is also a good idea, if possible, to present the fixed calibration locations in random order, to avoid any regular bias or expectation by the observer.

Calibration routines may be experimenter-controlled or system-controlled. An experimenter-controlled routine starts by announcing or showing the point to be observed. The experimenter then controls the length of sampling time at that location (e.g., by mouse click), to ensure a valid and reliable sample of x/y-observation locations. It is helpful, in this case, if the experimenter has an indication of sample-to-sample variance at each point, to know which, if any, points must be re-sampled. An automated procedure will randomly select a location to be viewed, then will sample the observer's point-of-regard until a variance criterion is achieved. The experimenter should, however, still be able to take control of the calibration in case of difficulties.

Maintenance of an eye tracking system is not usually a serious issue. Lenses, mirrors, and half-reflective surfaces should be kept clean and scratch-free. Cables must be kept clear of moving parts. Optics on head-mounted systems can be particularly delicate, so care must be taken when transporting, storing, and adjusting these headsets. In studies in which participants are moving, great care must be taken to keep cables free and clear.

How many participants are needed for eye tracking studies?

While the number of participants required is always domain-dependent, studies with an eye tracking methodology and within-participant manipulations have generally used relatively few participants, much like psychophysics or physiology studies. Table 1 summarizes the number of participants from several eye tracking studies, noting whether the experimental design was within or between participants, and providing a very short task description. Overall, designs have used about 6-30 participants. For within-participant designs, therefore, the CIF requirement of 8 participants per user type should generally be sufficient for eye tracking studies.

Table 1. Number of Participants from Recent Eye Tracking Studies

Number of Participants	Design*	Task	Reference
30	W	Searching variably-ordered lists	Crosby & Peterson (1991)
20	W	Searching menus for stated target items	Aaltonen *et al* (1998)
17	W	Search of four different web pages for specific information tagets	Cowan (2001)
11	W	Searching menu lists for target characters	Byrne *et al* (1999)
10	W	Matching icons of various sizes	Jacko *et al* (2000)
8	W	Searching alternate file tree representations	Pirolli *et al* (2001)
7	W	Using screens from Prodigy	Benel *et al* (1991)
25	B	Evaluation of a computer-delivered questionnaire	Redline & Lankford (2001)

*W: Within participant; B: Between group design.

4. RESEARCH NEEDS

The expanding field of applied eye tracking can provide a useful contribution to the field of usability evaluation. This chapter has provided information to aid in the cross-education of usability specialists on the basics of eye tracking, and to inform eye tracking specialists on the methods and requirements of usability evaluation. The graceful melding of these areas still, however, requires several advances in knowledge and practical methodology.

Usability specialists are under great time constraints during testing sessions, and more easily used tools are required to allow them to incorporate eye tracking into usability tests. Automated calibration procedures that could be executed between tasks would be helpful. These could operate as the usability specialist prepares for the next presented task. In addition, the eye tracker should be extremely quick to setup for a participant; for example, the tester might simply align a shape on a screen with the participant's pupil image.

Continued development of algorithms to compensate for head movement is recommended. Participants in a usability evaluation should not feel as if they are motion-constrained while using products. Also, better compensation for head motion during calibration might result in fewer participants who fail to calibrate to an eye tracker.

Improved eye tracking analysis tools are required to easily interpret data from studies where participants search and scroll across many screens or web pages. Currently, areas of interest must be tediously assigned and saved for each page; tools that would automate this process would enable the usability specialist to rapidly obtain a sequential order of visited areas on a task.

Convergence into a standard eye tracking protocol is desirable for usability testing, much like the NIST CIF standard for usability. Minimum standards, such as frequency of calibration, data sampling rates, and equipment specifications would make it easier to cross-interpret data from multiple studies. Certainly, more published reports of eye tracking in usability evaluations will be required before such a standard could be considered. Absolute, rather than relative benchmarks for eye tracking metrics would also aid cross-interpretation, and eventual convergence to a standard (Cowen, 2001).

More knowledge is required on the contribution of task factors to eye tracking-derived metrics. Little is known, for example, about how the density of a display or visibility of icons influences eye tracking results, and perceived usability. This is not a trivial issue, and has been the subject of visual search studies for many years.

The usability specialist is interested in whether software is above or below a stated usability threshold, and if below, how it may be improved. While there are well-accepted, standard usability metrics, little has been done to correlate eye tracking-derived measures to those. There is a strong need for research relating these two sets of measures.

Ultimately, eye tracking may become commonplace, as a secondary methodology for usability evaluation. This could help to drive down the price of eye tracking systems, and support an easier process of reducing eye tracking data from gazepoint samples to meaningful behavioral inferences.

5. REFERENCES

Aaltonen, A., Hyrskykari, A., and Raiha, K. (1998). 101 Spots, or How Do Users Read Menus? *Proceedings of ACM CHI 1998 Conference on Human Factors in Computing Systems 1998*. ACM Press, pp. 132-139.

Benel, D. C. R., Ottens, D. & Horst, R. (1991). Use of an Eyetracking System in the Usability Laboratory. *Proceedings of the Human Factors Society 35th Annual Meeting*. San Francisco, CA. 461-465.

Brookings, J.B., Wilson, G.F., and Swain, C.R. (1996). Psychophysiological Responses to Changes in Workload During Simulated Air Traffic Control. *Biological Psychology*, 42: 361-377.

Byrne, M.D., Anderson, J.R., Douglass, S., and Matessa, M. (1999). Eye Tracking the Visual Search of Click-Down Menus. *Proceedings of ACM CHI 1999 Conference on Human Factors in Computing Systems 1999*, pp. 402-409.

Card, S.K., Pirolli, P.,Van Der Wege, M., Morrison, J.B., Reeder, R.W., Schraedley, P.K., & Boshart, J . (2001). Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability Information Scent. *Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems 2001* p.498-505.

Carpenter, R. H. S. (1988). Movements of the Eyes. (2nd ed.). London: Pion.

Chin, J.P., Diehl, V.A., and Norman, K. (1988). Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. *Proceedings of ACM SIGCHI '88* (Washington DC), 213-218.

Cowen, L. (2001). *An Eye Movement Analysis of Web-Page Usability*, Unpublished MSc Thesis, Lancaster University, UK.

Crosby, M.E., and Peterson, W.W. (1991). Using Eye Movements to Classify Search Strategies, *Proceedings of the 35th Annual Meeting of the Human Factors Society*, 1476-1480.

Crowe, E.C., and Narayanan, N.H. (2000). Comparing Interfaces Based on What Users Watch and Do, *Proceedings of ACM/SIGCHI Eye Tracking Research & Applications Symposium*, pp. 29-36.

Dumas, J.S. (in press). User-Based Evaluation. In Jacko, J. & Sears, A. (Eds.), *Handbook of Human-Computer Interaction*, Mahwah, NJ: Lawrence Erlbaum, Inc.

Goldberg, J.H. (2000). Eye Movement-Based Interface Evaluation: What Can and Cannot be Assessed? *Proceedings of the IEA 2000/HFES 2000 Congress (44th Annual Meeting of the Human Factors and Ergonomics Society)*, Santa Monica: HFES, pp. 6/625 – 6/628.

Goldberg, J.H. and Kotval, X.P. (1998). Eye Movement-Based Evaluation of the Computer Interface. In Kumar, S.K. (Ed.), *Advances in Occupational Ergonomics and Safety*, Amsterdam: IOS Press, pp. 529-532.

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

Goldberg, J.H., Stimson, M.J., Lewenstein, M., Scott, N., and Wichansky, A.M. (in press). Eye Tracking in Web Search Tasks: Design Implications. *Proceedings of ACM/SIGCHI Eye Tracking Research & Applications Symposium 2002*, New Orleans, LA.

Goldberg, J.H., Probart, C.K., and Zak, R.E. (1999). Visual Search of Food Nutrition Labels. *Human Factors*, 41(3): 425-437.

Hoeks, B., and Levelt, W.J.M. (1993). Pupillary Dilation as a Measure of Attention: A Quantitative System Analysis. *Behavior Research Methods, Instruments & Computers*, 25(1): 16-26.

Jacko, J.A., Barreto, A.B., Chu, J.Y.M., Bautsch, H.S., Marmet, G.J., Scott, I,U., and Rosa, R.H. (2000). *Proceedings of ACM/SIGGRAPH Eye Tracking Research & Applications Symposium*, p. 112.

Jacob, R.J.K. (1995). Eye Tracking in Advanced Interface Design, in Barfeld, W., and Furness, T. (Eds.), *Advanced Interface Design and Virtual Environments*, Oxford Univ. Press, pp. 258-288.

Just, M.A., and Carpenter, P.A. (1976a). Eye Fixations and Cognitive Processes, *Cognitive Psychology*, 8: 441-480.

Just, M.A., and Carpenter, P.A. (1976b). The role of Eye-Fixation Research in Cognitive Psychology. *Behavior Research Methods & Instrumentation*, 8(2): 139-143.

Karn, K., Ellis, S., & Juliano, C. (2000). The Hunt for Usability: Tracking Eye Movements. *SIGCHI Bulletin*, November / December, p. 11. NY, Association of Computing Machinery.

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and Usage. In Jordan, P., Thomas, B., and Weerdmeester, B. (Eds), *Usability Evaluation in Industry*. London: Taylor & Francis.

Kotval, X.P., and Goldberg, J.H. (1998). Eye Movements and Interface Components Grouping: An Evaluation Method. *Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society.* Santa Monica: HFES, pp. 486-490.

Lankford, C. (2000). GazeTracker: Software Designed to Facilitate Eye Movement Analysis. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, November 6th-8th 2000, Palm Beach Gardens, FL, pp. 51-55.

Macleod, M., Bowden, R., Vevan, N., and Curson, I. (1997). The MUSIC Performance Measurement Method. *Behaviour and Information Technology* 16 (4-5), 279-293.

May, J.G., Kennedy, R.S., Williams, M.C., Dunlap, W.P., and Brannan, J.R. (1990). Eye Movement Indeces of Mental Workload. *Acta Psychologica*, 75: 75-89.

Mulligan, J.B. (1997). Image Processing for Improved Eye-Tracking Accuracy. *Behavior Research Methods, Instruments, & Computers*, 29:54-65.

Noton, D., and Stark, L. (1970). Scanpaths in Saccadic Eye Movements While Viewing and Recognizing Patterns. *Vision Research*, 11: 929-942.

Pirolli, P., Card, S. K., and Van der Wege, M. M. (2001). Visual Information Forgaing in a Focus + Context Visualization. *CHI Letters*, 3(1), 506-513.

Polson, P., Lewis, C., Rieman, J., & Olson, J. (1992). Cognitive Walkthroughs: A Method for Theory-Based Evaluation of User Interfaces. *International Journal of Man-Machine Studies,* 36, 741-773.

Potter, M.C. (1983). Representational Buffers: The Eye-Mind Hypothesis in Picture Perception, Reading, and Visual Search. In Rayner, K. (Ed.), *Eye Movements in Reading: Perceptual and Language Processes*. New York: Academic Press, pp. 413-437.

Rayner, K. (1978). Eye Movements in Reading and Information Processing. *Psychological Bulletin*, 85(3): 618-660.

Redline, C.D., and Lankford, C.P. (2001). Eye-Movement Analysis: A New Tool for Evaluating the Design of Visually Administered Instruments (Paper and Web). *Proceedings of American Association of Public Opinion Research*, Montreal, Canada, 5/01.

Sales, B.D., and Folkman, S. (2000). *Ethics in Research with Human Participants*. Washtington, DC: American Psychological Association.

Scholtz, J., Wichansky, A., Butler, K., Laskowski, S., and Morse, E. (in press). Quantifying Usability: The Industry Usability Reporting Project. *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society.* Santa Monica, CA: HFES.

Smith, B.A., J. Ho, W. Ark, S. Zhai (2000). Hand Eye Coordination Patterns in Target Selection. *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, November 6th-8th 2000, Palm Beach Gardens, FL, pp. 117-122.

Stern, J.A., Boyer, D., and Schroeder, D. (1994). Blink Rate: A Possible Measure of Fatigue. *Human Factors*, 36(2): 285-297.

The Poynter Institute (2000). Stanford University Poynter Project. *http://www.poynter.org/eyetrack2000/*.

Wichansky, A.M. (2000). Usability Testing in 2000 and Beyond. *Ergonomics*, 43(7), 998-1006.

Young, L.R., and Sheena, D. (1975). Survey of Eye Movement Recording Methods, *Behavior Research Methods & Instrumentation*, 7: 397-429.

Zellweger, P.T., Regli, S.H., Mackinlay, J.D., and Chang, B.W. (2001). The Impact of Fluid Documents on Reading and Browsing: An Observational Study. *Proceedings of ACM SIGCHI 2000 Conference on Human Factors in Computing Systems*, 249-256.