

Literature Survey:

Eye-tracking

and

User Performance Evaluation

Allen V.R. Harper
Department of Computer Science
City University of New York Graduate Center
May 16th, 2011

Table of Contents

1	Abstract.....	1
2	Introduction	2
3	HCI and Usability.....	7
4	Eye-tracking.....	10
5	Motivating Example	17
6	Paper Selection and Evaluation Criteria.....	19
7	Paper Critiques	24
7.1	ATC Study.....	25
7.2	IDE Study.....	32
7.3	PILOT Study	38
7.4	LINE Study	44
7.5	NEWS Study.....	49
8	Summary of Findings	55
9	Conclusions	58
10	Future Directions	60
11	References.....	62

List of Figures

FIGURE 1 SCAN PATH SUPERIMPOSED ON AOI	14
FIGURE 2 EXAMPLE HEAT MAP	15
FIGURE 3 MOCKUP OF AN EDUCATIONAL LEARNING SYSTEM INTERFACE	17
FIGURE 4 ATC SIMULATOR INTERFACE	26
FIGURE 5 A) TEXT UI B) COLOR UI C) PFT RESULTS	30
FIGURE 6 USER INTERFACE FOR IDE PAPER	33
FIGURE 7 ORIGINAL AGGREGATED DATA FOR IDE STUDY	35
FIGURE 8 SEGMENTED PRESENTATION OF PFT	36
FIGURE 9 LANDING SIMULATOR UI FOR PILOT STUDY	39
FIGURE 10 PILOT LANDINGS A) EXPERTS (ON LEFT) B) NOVICES (ON RIGHT)	41
FIGURE 11 A) EXPERT (LEFT) B) NOVICE (RIGHT) SAMPLE SCAN PATHS	42
FIGURE 12 TOTAL FIXATIONS/AVERAGE DWELL TIME EXPERTS AND NOVICES	43
FIGURE 13 FIXATIONS AND PERFORMANCE	43
FIGURE 14 EXAMPLE OF A MODIFIED SCAN PATH IN THE LINE STUDY	47
FIGURE 15 THREE VERSIONS OF THE NEWS STUDY UI	50
FIGURE 16 PERFORMANCE RESULTS FOR NEWS STUDY	52
FIGURE 17 PFT RESULTS FOR NEWS STUDY	53

1 Abstract

Evaluating how people perform tasks while they interact with user interfaces is a well-studied problem within Human-Computer Interaction (HCI) research. The traditional goal of such usability evaluations has been the determination of the effectiveness (error rate) and efficiency (speed to task completion) of task performance using a particular interface.

During the past decade a technique which tracks the user's eye movements has been employed as a complimentary method to assist in evaluating task performance. By determining the coordinates of where a person is focusing while they perform a task, eye-tracking produces a much more fine-grained record of **how** a task is completed than is possible with the more traditional HCI methods (e.g., surveys, Talk-Aloud Protocols).

Recent directions within eye-tracking research have focused on novel uses of the data collected in an eye-tracking experiment. These newer methods have focused on the correlation of patterns in the eye-tracking data related to the qualities of the user's interactions. One possibility, and the focus of this literature survey, is the connection between eye-tracking data and **how well** users perform a task. Establishing such a correlation is significant in that it would allow for the design of automated computer systems capable of detecting when users are in need of assistance.

Unfortunately, the eye-tracking literature does not yet contain complete studies with the explicit goal of measuring this connection between eye-tracking and performance. However, many examples of eye-tracking studies do exist which provide implicit, preliminary evidence that such a connection does in fact exist. The purpose of this paper is to evaluate the recent eye-

tracking literature with the goal of designing a framework within which future eye-tracking experiments could be conducted which addresses the connection between eye-tracking and performance.

2 Introduction

Human interactions with computers can be considered from the perspective of two information processors interacting across a user interface (UI) (Jacob 1995). As such, the study of human behaviors related to these interactions is of particular interest within the field of Human-Computer Interaction (HCI) as they can provide insight into human performance. One class of behavior that is of particular interest to HCI research is eye movements. The reason for this interest is that many groups of computer users can receive as much as 80% of their perceptual input during interactions (Cuddihy, Guan et al. 2005). Therefore, techniques for studying eye movement behavior have been considered an effective means of characterizing human-computer interactions.

The significance of eye movements to usability testing is that they represent an overt, observable and therefore, measurable quantity associated with the user interaction which are assumed to have a predictable relationship with the covert cognitive processes associated with visual attention which in themselves are currently not thoroughly understood. As a result of this relationship, eye-tracking measures, such as total number of fixations, gaze durations, and scan paths (defined in Section 4) can provide detailed information as to *how* users perform tasks—information that would be difficult (if not impossible) to gather using other HCI methods. As Kowler explains:

Eye movements and attention are assumed to serve useful purposes connected to the visual task, an assumption that has fueled decades of efforts to use eye movements to study how people search, read, study pictures of scenes, or carry out all manner of visually-guided actions involving reaching, pointing, manipulating objects, walking, or driving (Kowler 2006).

These efforts to develop techniques to record and interpret the eye movements of experimental subjects are detailed in many recent surveys within this specialized area of the HCI usability literature (Rayner 1998; Duchowski 2002; Jacob and Karn 2003; Poole and Ball 2005). Due to the proliferation of the use of eye-tracking (ET) across many disciplines this literature can appear quite fragmented; However, one important theme running throughout this literature is the significant contribution to the understanding of *how* users perform tasks which is provided by the analysis of eye movements. Leveraging eye-tracking in this fashion has led to improvements in the understanding of user behavior during interactions as well as to improvements of traditional HCI measures such as Think-Aloud protocols (Cooke and Cuddihy 2005; Guan, Lee et al. 2006; Eger, Ball et al. 2007). In our survey, we will refer to this explanatory use of eye-tracking data as the HCI/Eye-Tracking or **HET** perspective. The goal of the HET perspective is to compliment HCI methods which attempt to explain how people go about performing tasks.

While the HET perspective represents the most common use of eye-tracking in HCI experiments, it is not without its weaknesses. It has been argued that HET has numerous limitations which have created significant barriers to its ease of use as an HCI research tool (Jacob and Karn 2003). For example, the lack of straightforward methods for working with the larger more complex experimental data sets typically generated in an eye-tracking experiment—some eye-tracking experiments have collected data for approximately thirty minutes (Helleberg and Wickens 2003); the lack of standard and well-understood reporting practices and

interpretation of eye-tracking results paralleling the more traditional HCI methods found in usability research such as reaction time (RT), speed of task execution and user error rates (Jacob and Karn 2003) (Medina, Cuddihy et al. 2008). Most important to the goals of our survey is the inability to consistently relate eye-tracking measures to user outcomes such as task-performance. Taken together, all of these limitations have made it difficult for eye-tracking research to make cross-study comparisons and therefore generalize their eye-tracking findings—a theme which will be reiterated throughout this survey.

Our survey proposes to address the drawbacks of the HET approach by reviewing the literature from a slightly different perspective with the goal of highlighting the elements of an alternative framework. To do this, we will shift away from using eye-tracking data to explain *how* users perform tasks to a new focus on measuring *how well* the users perform the task. In order to make it clear which perspective we are using to discuss a particular paper we will be referring to our alternative framework as the Eye-tracking /Performance Connection or *EPC* perspective. To clarify the difference between the two perspectives, researchers who use eye-tracking data in a HET manner try to use the recording of eye movement during an experiment to give an informal explanation of what the person was doing during the experiment – specifically, how they were accomplishing the task. A researcher who uses eye-tracking data in an EPC manner would attempt to use a recording of eye-movement during an experiment to determine how well a user was accomplishing some tasks – specifically, what is the user’s likely success as some task based on an analysis of the eye-tracking data. As we stated in Section 2, the EPC approach to using eye-tracking data is a novel proposal in this literature survey for which there is only a sparse research literature that connects to this topic.

Shifting the emphasis from HET to EPC has potential benefits for both usability evaluation as well as the design of computer systems in general. By establishing a predictable relationship between eye movements and user performance, it is feasible that eye-tracker data could function as a proxy variable for task performance. In terms of usability evaluation, considerable reduction in both time and effort involved in conducting usability studies could be achieved if user performance could be determined via the automated use of modern eye-tracking equipment as opposed to those employed in more traditional HCI frameworks (e.g., post-trial interviews). A second benefit of employing eye-tracking from the EPC perspective is that it allows for the inclusion of users with permanent disability (e.g., Deaf users using sign language) where traditional evaluation techniques (Talk-Aloud protocols) would interfere with task execution (due to hand movements) (Roberts and Fels 2006). A third benefit of EPC approach is the potential to provide real-time monitoring of user performance whereby eye-tracking data is incorporated into *affective* UIs which can both detect and respond to the current state of the user (e.g., frustration) by providing intelligent feedback to the user (Sibert, Gokturk et al. 2000; Ratwani, McCurry et al. 2008) (Hardoon and Pasupa 2010). For example, in Sibert et al. (2000) a computer system designed for remedial reading practice detects via eye movements when users are struggling with words and provides assistance.

Currently, within eye-tracking research, establishing a direct connection between eye-tracking measures and user performance is not an explicit research goal. This is a novel research direction that we are actively considering at CUNY, and there are no papers in the literature that are explicitly searching for concrete connections between eye-movement and user success at tasks during HCI usability experiments. However, within this body of literature we have found

evidence for EPC within research papers that had different stated research goals; most of these papers were conducted within the more traditional HET framework (with an eye-tracker being included in an experiment with a person interacting with a computer for the purpose of providing extra data that the researcher might use to characterize or explain user actions after the fact).

This survey will argue that in order to scientifically establish that a connection exists between eye-movements and users' task-performance it will be necessary to conduct carefully designed *EPC verification* experiments. Namely, an EPC verification experiment is an eye-tracking study in which subjects attempt to perform a task with a given user interface with varying levels of success. Following the collection of both performance and eye-tracking data, researchers conduct a statistical analysis in search of correlations between eye-movement metrics and subject's performance scores. These experiments would tightly control for such factors as subject selection, the user interface, the visual content displayed to users, as well as the task type and difficulty to be performed. The importance of providing stringent controls on these factors is the great sensitivity of eye movements to almost any change in an experimental design. By controlling for these factors it becomes possible to establish predictable associations between eye-tracking metrics and user performance – through some form of statistical analysis of the eye-movements of users in the experiment and their performance on some task.

Problematically, as discussed above, EPC-style verification experiments have not been an explicit research goal in the eye-tracking literature. Secondly, locating eye-tracking studies which have controlled for the experimental factors mentioned above (i.e., good subject selection, etc.) are also not well represented in the HCI eye-tracking literature. As a result of this wide

variation in the quality of eye-tracking experiments found in the eye-tracking literature, this survey proposes to select a subset of what we believe to be exceptional state-of-the-art HET experiments which are constrained in such a way that they will provide a coherent (if only preliminary) view of the appropriate experimental guidelines for conducting EPC verification experiments in the future. So, in our survey of the literature we will analyze papers from the perspective of whether some of the results found in the papers surveyed suggest that there exists a definable relationship between eye-movements and user task-performance for some user-interface. While the original research goal of the papers included in our survey was not to establish such a connection, we will reanalyze the papers with this new perspective in mind.

To further define the framework necessary to carry out the goal this survey, short discussions are included below which will provide the appropriate background information. Section 3 will discuss a particular HCI framework with respect to experimental procedures that this survey will use as a measure in selecting papers. Section 4 will perform a similar task for the area of eye-tracking by describing pertinent aspects of the technology and by normalizing and defining the terminology found in the literature. Section 5 will describe a particular type of computer application that will serve as a motivating example and reference point. Section 6 will cover the criteria used to both select and evaluate the papers as well as a brief description of the process followed for locating the papers in this survey. Section 7 contains the analyses of the selected papers and Section 8 presents the summary of our findings. Finally, Sections 9 and 10 present the conclusions of this survey and discuss future directions of this research, respectively.

3 HCI and Usability

The field of Human-Computer Interaction (HCI) employs a variety of definitions and concepts related to usability, usability experiments, and usability measures (Hornbaek 2006). In an attempt to provide a framework for comparison and discussion of the experimental designs used within the eye-tracking experiments discussed in this survey, working definitions will be provided here.

Earlier, we discussed Jacob's model of human-computer interaction of two information processors communicating across an interface. The following definition of HCI provided by the Special Interest Group on Human Computer Interaction (SIGCHI) is also useful to consider:

Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them (Hewett 1992).

In this context, evaluation of usability typically refers to the goal of determining the “effectiveness, efficiency, and satisfaction with which specified users can achieve goals in particular environments” (ISO 1998). Within the field of HCI there exists a variety of evaluation methodologies which have been developed over the years (e.g., Ethnographic, Expert, and Heuristic). However, in the context of this survey the term *usability evaluation* will mean the use of an experimental approach with the aim of providing empirical results which lend support to a particular hypothesis for some user-interface (UI). In practical terms, measuring usability of a system is done by assessing the functionality of the UI, how the UI affects user performance characteristics, and the identification of specific problems with the system which could lead to diminished usability. To perform such measurements requires the collection of experimental data associated with user interactions via the interface.

Measures of usability typically fall within two categories: effectiveness (error rate) and efficiency (speed of task completion). In order to measure the effectiveness and efficiency of a UI it is necessary to operationalize these concepts into quantifiable experimental variables. Typically, this is done by creating experimental variables derived from performance measures (e.g., keystrokes per minute, errors per task).

In the case of effectiveness, it is customary to use performance measures that relate to the accuracy and completeness with which user tasks can be accomplished. In the literature, these measures are typically reported as either the raw number or percentage of correct actions that the user performed. In a particular instance, this might take the form of the number of correct answers, the number of incorrect actions (error rate), mouse clicks within a defined area, or number of items recalled or the number of sub-parts completed of a multi-part task.

Efficiency, on the other hand, is typically reported as a ratio of the effectiveness achieved as a fraction of the resources used. In this regard, resources might include physical energy, mental difficulty, money, or time. Time-based measures are very common in HCI studies; the “time to complete a task” is reported frequently. However, other time-based measures such as “time until the first occurrence of an event of interest” or “input rates” (keystrokes per minute) are also commonly reported measures. In the analysis of UIs, efficiency might also be measured in terms of the patterns of UI elements used – with some patterns being observed to lead to greater success.

While the discussion so far has mentioned the inclusion of a user task in a usability evaluation it needs to be stressed that the *type* of task is important with regard to capturing eye

movements. In particular, the term *task* in the context of our survey means a directed and intentional task requiring significant attentional resources for its successful completion. This definition therefore precludes tasks (e.g., picture viewing) which do not force the user into keeping pace with changes brought about on the user interface. As this survey will describe, the inclusion of a sufficiently attention-requiring task within an eye-tracking usability experiment provides an important structural element, which, when absent, leads to highly variable and less reliable experimental data.

Summing up, from the point of view of our survey, HCI is an experimental discipline which relies upon the collection of data from human subjects as they perform tasks with the aid of a computer system which they interact with via a user interface. This point of view implies that the quality of the data collected (e.g., eye-tracking and performance measures) is strongly dependent upon the conditions under which the experiment is conducted and therefore in order to obtain useful data from an experiment, it is important that significant effort goes into the design of an HCI experiment. From the point of view of our survey, experimental factors such as how subjects are chosen, the difficulty of the task to be performed, the complexity of the user interface and the visual content presented to the subjects are all important factors which need to be carefully examined prior to conducting actual experiments. As we will see in the next section of the survey, controlling for these experimental factors becomes an even bigger issue when the data being collected is related to human eye movements.

4 Eye-tracking

This section of the survey will provide an overview of the fundamentals of how eye-tracking works and how it is currently employed within HCI usability research. Common eye-tracking terms will be defined and the naming of eye-tracking terms will be standardized as the literature often employs various definitions.

Eye-tracking is a method of recording the geometric coordinates of a user's point of focus upon a visual stimulus. Numerous techniques have been developed to accomplish this measurement—all varying in their precision, invasiveness, and restrictiveness on the movement of experimental subjects. For example, “electro-oculography” (EOG) measures changes in electrical potentials surrounding subject's eyes via the attachment of electrodes to the skin surrounding the eye. The “magnetic search coil” technique places a specialized magnetic contact lens in the subject's eye while movements are detected by the deformation of a magnetic field generated by a cubic apparatus surrounding the subject's head. For cases where great precision is required, the “Dual-Purkinje Image tracker” is employed. This technique relies upon the measurement of changes in the position of reflected infrared light from the crystalline surfaces of eye structures (Hammoud and Mulligan 2008).

While the precision of some of these techniques is needed in various branches of research which employ eye-tracking (e.g., psychology of reading) the area of HCI research relating to the evaluation of user interfaces has opted to sacrifice some precision in favor of a less invasive and restrictive technique known as video-based corneal reflection (VCR). In VCR systems cameras unobtrusively attached to the video display capture images of the exterior of the eye and record the location of prominent eye structures and the reflection of a single infrared light source. Image

processing software then monitors the relationship of these two artifacts and calculates the coordinates of focus. Desktop mounted VCR hardware place few restrictions on the movement of subjects and provide a comfortable user experience.

All of these techniques used in the tracking of eye movements are directed towards the goal of determining the screen coordinates of the point where the eye is focused upon; therefore, it is important to understand the relationship between the mechanics of human vision and the types of eye-tracking measurements that can be made. While the human field of vision covers only about 200 degrees, the entire field is not rendered in the same degree of acuity. This results from the fact that the anatomy of the human eye only allows for a small portion of the field of vision to be kept in sharp focus. In particular, only a small structure at the rear of the retinal wall (fovea) has the necessary density of receptors to capture sufficient information in order to present the visual processing centers of the brain with a detailed image. The remainder of the visual field (parafoveal (2-5 degrees) and peripheral) is not in clear focus. Thus, of the 200 degrees of visual field approximately 1-2 degrees (a region about the size of a thumbnail at arm's length) is in clear focus at any given time (Richardson and Spivey 2004).

The human eye compensates for this inability to maintain the complete field of vision in detailed focus by being in continual motion. By continually sampling the visual scene via the fovea the visual processing regions of the brain are able to assemble a complete field of vision and thereby maintain the illusion that the complete field of vision is in sharp focus (Richardson and Spivey 2004).

Two primary actions are associated with this continual movement of the eye—*fixations*

and the *saccades*. *Fixations* are periods of decreased eye movement (not necessarily the complete lack of motion) which last between 200-250ms. During this time frame, the focal point is constrained within a space of approximately 1 visual degree. *Fixations* are used by the eye to focus on new targets as well as extract finer details from a particular region of the scene. *Saccades*, on the other hand, are ballistic movements reaching rotational speeds of 500 degrees per second during which the eye repositions itself on a new target of interest within the visual scene.

Eye-trackers typically employ algorithms to determine if the eye is currently fixating or in motion and will record this information along with the screen coordinates. In addition to the fixation data, an eye-tracking experiment also includes the definition of *area-of-interest* (AOI). AOIs are typically defined *a priori* and segment the user-interface into discrete sub-regions of interest to the research question.

Based upon the raw fixation data a number of eye-tracking measures can be derived. Jacob and Karn (2003) in their survey of numerous eye-tracking usability experiments report the use of such measures as total number of fixations, fixations per AOI, fixations per second, fixation duration, fixations per AOI, density of fixations per AOI as well as the probability of fixation on an AOI. Obviously this list is not exhaustive and indicates the creativity of researchers in discovering explanatory patterns within the eye movement data.

Scan path, is a third eye-tracking measure commonly reported in the literature (See Figure 1). By joining all fixation points with line segments a path of visits to AOIs within the UI is generated. Scan path data also provides a good example of the level of noise that is found in

eye-tracking data and the need for strict experimental control.

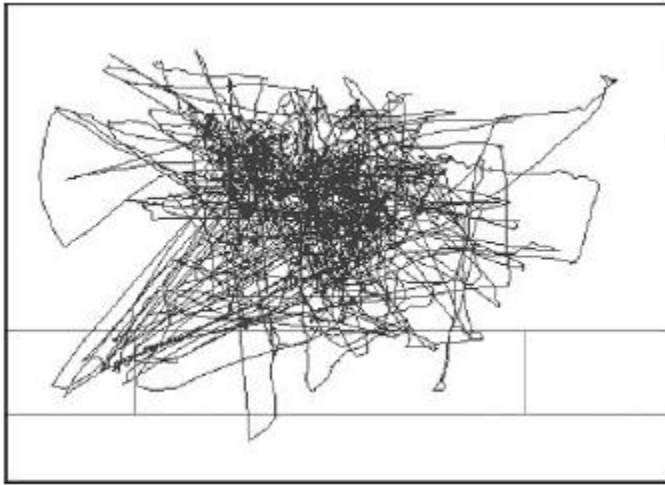


Figure 1 Scan path superimposed on AOIs (Josephson and Holmes 2006)

Another commonly reported measure refers to the grouping of a number of distinct fixations—all temporally related to one another—that fall within an AOI. Numerous terms within the literature have been applied to this description (e.g., *dwell*); however in this survey the term *gaze* will be applied.

Based upon this definition of gaze numerous other eye movement measures have been derived and reported in the literature. Again, Jacob and Karn (2003) have documented the use gaze rate, mean gaze duration per AOI, and gaze % per AOI. Finally, the concept of gaze can be used in an analogous fashion to scan path with individual fixations being replaced by gaze on a particular AOI. As gaze is shifted from AOI to AOI a *transition* is recorded with the total list of AOIs forming a string (A1, A2, A1, A4, etc.). This string representation of visited AOIs has led to the development of techniques for analyzing eye-tracking data based upon the algorithmic analysis of strings (similar to those being applied in the area of Bioinformatics).

Another popular representation of gaze duration which appears frequently in eye-tracking studies is called a *heat map* (see Figure 2) in which red colors are overlaid upon the area of the UI that receive longer total gaze durations while blue colors represent areas receiving less visual attention.

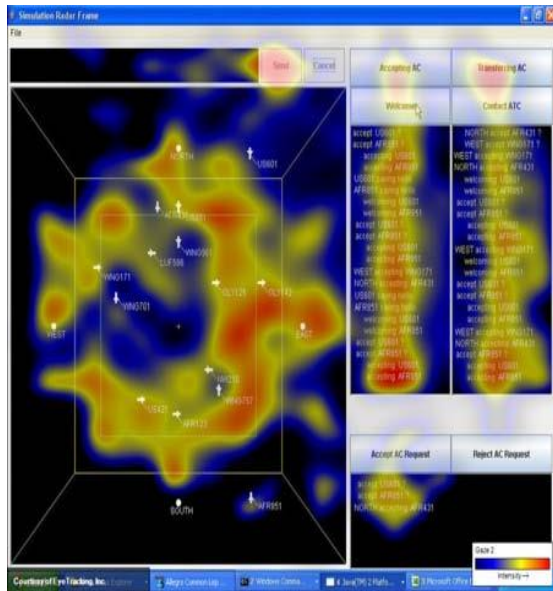


Figure 2 Example heat map (Bartels and Marshall 2006)

This representation is valuable for gaining a quick understanding of AOIs which were important to users; however, by lacking any temporal aspect of the gaze data and summing over the length of the experiment this technique disallows the timing aspect – a feature that is brought forward in the analysis of transitions from AOI to AOI a point that will be highlighted below in the analysis of one paper in particular (Bednarik and Tukiainen 2008).

The ease of use of the modern computer-based VCR systems has led to its proliferation in HCI research. Focusing only on systems which contain user interfaces the domain can be partitioned into two areas–Interactive and Diagnostic (Duchowski 2002). On the Interactive side

are the eye-tracking applications that use eye movements as HCI input for either disabled or “hands busy” applications (Selective), interfaces that are altered by user’s eye movement patterns (Gaze-Contingent), as well as the newest field—Affective interfaces—which detect user state (e.g., frustration during a search task). Duchowski’s survey, however, is primarily concerned with diagnostic uses of eye-tracking in which users will be presented with visual stimuli on a video display while eye movements are recorded for the purpose of determining characteristics of the user interaction.

In summary, diagnostic eye-tracking evaluations are typically performed in laboratory settings with selected users performing tasks via a user interface. The preferred eye-tracking method is the video-based corneal reflection technique. Given the greater length of eye-tracking experiments (upwards of 10 minutes) and the sensitivity of human eye movements, data sets collected during eye-tracking experiments are typically larger and noisier than found in more traditional HCI experiments. As a result, eye-tracking researchers have developed methods to handle this issue which include dividing the user interface into important regions (areas of interest) and tallying the eye-tracking metrics per AOI. Thus, we typically see reports of fixation counts or gaze durations per AOI. Another summarizing technique for the eye-tracking data is the use of heat maps which represent the intensity of fixations on the user interface. Finally, scan paths are used to summarize the total path traveled by the eye during an experiment.

5 Motivating Example

As the previous discussions indicate, eye-tracking is widely used in many research domains. Considering this fact, the decision was made to narrow the choice of eye-tracking studies included in our survey to applications of eye-tracking which had a significant overlap with a particular application domain: educational learning systems.

Figure 3 represents a prototype of a UI typical of those seen in this emerging area. In many cases, such systems are deployed as accommodations for special user groups within a classroom setting. This particular UI design is an amalgam of UI styles currently being tested with Deaf users (Hughes and Robinson 2007; Cavender, Bigham et al. 2009)—a user group of potential interest to the future directions of this research.

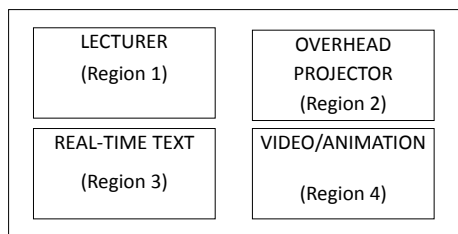


Figure 3 Mockup of an Educational Learning System interface

Figure 3 displays four regions which contain different types of information important to the user. Regions 1 and 2 present academic lecture information that all users might find in a typical educational environment—video of the lecturer and slides. Regions 3 and 4 present two specialized panels. On the lower left is a panel displaying a transcript of the lecturer’s speech (perhaps generated in real-time via speech recognition software) and in the lower right a sign

language representation of the lecture either via a live translator or an off-the-shelf animation package typically transliterating the lecture text into animation characters.

One significant feature of this user-interface which makes it important for our survey is the high degree of “visual dispersion.” That is, information is not contained within a single UI element, but instead is located in discrete UI elements across the interface. This type of interface will require the subject to make eye movements which transition from region to region. A second feature of importance for our survey is the dynamic nature of the underlying task which requires subjects to keep pace with the stream of information on the user interface. There is ample evidence in the literature to indicate that such tasks generate both speed stress and foveal loads which decrease subject’s usable visible field and therefore degrade the ability to notice events occurring in the periphery (such as slide changes)—an impact which leads to a decrease in task performance (Williams 1985) .

This combination of visual dispersion and foveal load that is found in applications of this type lends itself to the study of connections between eye movement behaviors and human performance. The reason for this is that performance is tightly coupled with subjects making frequent transitions and where they transition to impacts their performance. In such a scenario, subjects do not have the freedom to gaze at UI elements which are not directly involved in task performance—and as a result the data collected by the eye tracker is itself also tightly coupled with subject’s performance. As we will see in the paper discussions below, there is a complex interplay between the user interface and task specification on the one hand and human performance and eye movement behavior on the other.

6 Paper Selection and Evaluation Criteria

One of the challenges we faced in performing our survey of the eye-tracking literature was the selection of a coherent set of papers. Unlike more traditional research topics with an already circumscribed area of focus, in our case we also had the task of constructing the boundaries for our research area. This explains the effort that we have put into clarifying what we mean by an HCI experiment, pointing out what experimental conditions we believe might impact the data in an eye-tracking study, and also defining a motivating example. The purpose of all of this work was to establish a filter with which the eye-tracking literature could be screened in order to obtain the desired set of papers.

A second challenge faced by our survey was actually locating the papers which fit our definition. While it might appear that our survey only relied upon the major sources of eye-tracking and HCI research (namely the ACM sponsored Eye Tracking Research and Applications (ETRA) and SIGCHI) in fact many times the number of referenced papers in this survey were reviewed and rejected for a variety of reasons. Some of the difficulty in locating papers is attributable to the simple lack of a standard search term for eye-tracking. “Gaze Tracking”, “Eye Movement Study”, “Scan Behavior” are all typical examples of the type of search terms that needed to be employed. Once papers had been located that contained this search term it was then necessary to determine if the paper was in fact referring to tracking eye movement or was instead considering other eye behaviors such as pupil dilation. The end result was that the only sure way to cover the literature was to repeatedly follow the citation trails of well-cited papers and surveys.

Once a large pool of papers had been located in the literature, the next step was to apply the criteria as developed in Sections 1-5 of our survey. The important distinction between the focus of previous HCI research that uses eye-tracking and our current focus on the issue of EPC is that prior studies used eye-tracking data to characterize *how* users perform their tasks. From an EPC perspective, the question being explored is whether or not a connection between eye movement and user performance can be identified in the user data under the specific conditions presented above. We want to know whether eye-tracking data can be used to determine or predict *how well* the users have performed. In addition to identifying experimental conditions in which EPC is observed, we are also interested in analyzing experiments in which it is not observed. We want to understand whether the lack of such a connection can be attributed to the absence of certain experimental conditions.

The following discussions will attempt to separate out the cross-cutting issues of experimental design and eye-tracking methodology and assess their separate and combined impact on the connection between eye movement behaviors and user performance. First, the selected papers should all have a strong HCI and eye-tracking foundation. Specifically, the papers should adhere to well established experimental practices in both HCI and eye-tracking research as described in Sections 3 and 4 above. Second, papers that include UIs with multiple-representations of information, a characteristic of the motivating example described in Section 5, will also be included. Finally, papers that also include evidence of EPC will be highly valued for inclusion in this survey.

The papers appearing in this survey have been drawn from such diverse areas as Cognitive Science, Software Engineering, and Aviation Psychology – to list a few. However, it

is important to note that the area of research that is of interest to this survey, namely EPC, is in most instances represented only implicitly in the literature. While the literature contains numerous HET studies which include both eye-tracking and performance measures, it is rarely the case that the connection between eye movements and performance is analyzed directly. Therefore, besides reviewing the selected papers on their original intent as examples of HET experiments, this survey can be seen as a *meta-analysis* of the included papers with the potential of extracting experimental features that would be important in conducting such experiments in the future with the goal of verifying the existence of EPC. Therefore, the papers are being primarily critiqued on their adherence to research principles which will allow for the identification of guidelines to conduct future experiments successfully.

While we are primarily analyzing these papers to determine if they shed any light on how to identify connections between eye-tracking data and user performance on tasks in HCI usability experiments, we can still critique the papers on how well they followed established research guidelines on HCI usability studies or eye-tracker data collection experiments. The reason why this is important is that the researchers in these studies must get the HCI and eye-tracking aspects of their research correct or else there is no chance that their data is good enough to enable us to perform our meta-analysis to look for EPC in their results. For example, if not following HCI experimental practices has a negative impact on eye-tracking results, this survey will point this out. These type of critiques fall within the original intent of all the included studies.

Given the contents of the previous discussions it is possible now to list the characteristics of the papers that will be preferred by this survey. From the HCI perspective, a preference will be shown for papers that adhere to the following principles:

- Contain clear and testable hypotheses
- Use an appropriate experimental design (Within vs. Between groups)
- Have clear selection criteria for participant groupings (Expert vs. Novice)
- Use of replicable experimental procedures (Researcher scripts)
- Include directed and intentional user tasks that are non-trivial
- Include the use of task-performance measures
- Provide statistical analyses of the results

From an eye-tracking research perspective, papers will be preferred which display the following characteristics:

- The definition of appropriate AOIs for the user-interface and the task
- Users are exposed to identical visual stimuli across experimental conditions. The user-interface layout should be the same, and the visual content presented in different experiments should also be similar (e.g., users may tend to look at horizon lines and faces in images; so, visual stimuli in different experimental conditions should attempt to use equivalent imagery)
- Use of an intentional task that requires the use of eye-gaze in order to be successful (e.g., a non-intentional task may be one in which the user browses

pictures without a stated goal to achieve and a non-eye-gaze-dependent task would be one that could be accomplished regardless of where the user looks)

We are most interested in identifying evidence of EPC within contexts that have significant overlap with the characteristics of the motivating example. Thus, we would also prefer to discuss papers in this survey that discuss an eye-tracking experiment with an interface that bears some similarity to the “Motivating Example” in Section 5. From this perspective, we would prefer papers that have the following characteristics:

- User-interfaces with multiple regions of the screen that contain information that the user needs to complete a task.
- The user must move their eye-gaze between these regions in order to continuously monitor these regions and gather the needed information to be successful at the task.
- The information content of the user-interface is dynamic and requires the user to keep pace with it.

Finally, from an EPC perspective, in order for us to be able to identify papers in which there is a measured connection between user performance and eye-tracking data, we prefer papers with the following characteristic:

- The paper must present their results in such a way that we can see a relationship between the users’ performance on a task and some metrics that record their eye-tracking behavior. This may be presented in a graph or a table that the authors include in the paper, or we

may be able to infer this relationship in an indirect way by comparing eye-tracking data presented for novice vs. expert users. (As discussed above, identifying an EPC connection is rarely what the authors were originally intended when they conducted their research; so, we are performing a meta-analysis on their presented results.)

7 Paper Critiques

In the following discussions, the survey will compare and critique five papers which bring forward the ideas presented in the previous sections of this survey. To help manage the paper discussions each paper has been assigned a descriptive nickname based upon a prominent feature of the paper. The papers will be discussed in the following order:

- **ATC** (air traffic controller) study: Bartels et al. (2006)
- **IDE** (integrated development environment) study: Bednarik et al. (2008)
- **PILOT** study: Kasarskis et al. (2001)
- **LINE** study: Uwano et al. (2006)
- **NEWS** study: Josephson et al. (2006)

7.1 ATC Study

The ATC (air traffic controller) study follows a familiar trend in the HCI literature in that a previous experiment is replicated with the addition of an eye-tracking component – with the goal of gaining further understanding of how human subjects perform tasks beyond the level of detail of typical HCI experimental methods (e.g., post-trial interviews). In the original study, changes in performance were observed under varying conditions of user interface design and the level of difficulty in the task; however, the relationship between how subjects interacted with the particular interfaces and why this impacted performance was not clearly understood. By including an eye-tracking component in the current ATC study, the authors hoped to solve this issue.

With the inclusion of the eye-tracking component, the ATC study becomes of interest to our survey because it allows us to perform the meta-analysis we described in Section 6. The ATC study is a particularly good example of the type of eye-tracking experiment of interest to us due to its strong similarities with what we have defined as a *verification* experiment, which can identify if there are connections between task-performance and eye-tracking measures.

The experimental design of the ATC study consisted of human subjects ($n=14$), selected based upon their expertise in a similar visual task (video games), ranging in age from 21-35 years of age and including 12 males and 2 females. The task was a simulated air-traffic control task involving the management of multiple aircraft traversing the airspace (represented by the left portion of the UI displayed in Figure 4) while simultaneously monitoring for and responding to messages displayed in separate UI panels on the right portion of the UI.



Figure 4 ATC simulator interface (Bartels and Marshall 2006).

The task was composed of a set of five (5) distinct subtasks for the correct handling of each aircraft. The subtasks had to be completed in a specified order each requiring the subjects to read and respond to multiple request-response message pairs.

The experiment was conducted over a three-day period allowing subjects to familiarize themselves the task as well as for researchers to properly calibrate the eye-tracking equipment. The experimental design of the eye-tracking component divided the ATC simulator UI into seventeen (17) AOIs over which proportional fixation time (PFT) per AOI and transitions between AOIs were recorded.

Three levels of task difficulty were generated by decreasing the total time allotted for each experimental trial in order to handle a fixed number of aircraft and performance was measured as a function of accumulated penalty points which accrued for failing to correctly respond to the aircraft during each of the subtasks.

Besides controlling for task difficulty, the ATC study also presented subjects with two versions of the simulator interface. The first, a text-based UI, indicated changes in aircraft status

solely based on text messages. A second version of the UI was identical to the text-based version except that a color-coding scheme was added in which each color corresponded to a particular aircraft status message. This arrangement allowed subjects to monitor aircraft status without the need of reading the incoming message in the message panels of the simulator UI.

Given this description of the ATC study, we can identify many of the experimental criteria which our survey has listed as important elements of a verification experiment. In the ATC study these would include: the careful handling of experimental design features such as subject selection, the use of experimental procedures so that all subjects received similar treatment, and the unusual amount of time devoted to subject preparation. By including two days of familiarization and practice trials subjects were able to gain proficiency in task performance; this should contribute to the quality of both performance and eye-tracking data. As a result, this careful planning should lend confidence to the results as they apply to our survey's goal of identifying linkages between performance and eye-tracking measures.

Besides the handling of these procedural issues of the experimental design, the inclusion of a complex, goal-oriented, and time-constrained task is also of significance. Such tasks have been shown to create the necessary foveal load and speed stress capable of narrowing subject's usable field of view—an effect often referred to as *tunneling*. The importance of inducing this effect in eye-tracking studies is that under this condition, subjects find it more challenging to move their visual attention into areas of interest that are not directly related to the performance of the task. By controlling visual attention in this fashion, subjects are forced into more economical

patterns of eye movements and as a consequence eye-tracking data will contain less “noise”—i.e., AOIs will not accumulate PFT or fixation counts unrelated to task performance.

A second important characteristic of the task used in the ATC study was the manipulation of difficulty across three (3) discrete levels. Since the goal of our survey is the identification of connections between eye-movement and performance, it will be essential that there is a range of performance data in order to allow for the correlation of eye-tracking and performance metrics.

Additionally, as we will see in this and other papers presented in our survey, alterations to either the visual content (i.e., information presented within the UI) or user interface itself can have dramatic and sometimes unexpected effects on eye movement behaviors. Therefore, in eye-tracking experiments, it is preferable if both the visual content and the UI regions remain unchanged throughout the experiment. However, in the case where the UI itself is the independent variable and will therefore be presented to subjects in different formats (as was done here in the ATC study) it is that all exposure to the visual stimuli be handled in a controlled manner. Therefore, it was significant that in the ATC study exposure to the two UI types was handled in a balanced fashion.

Concluding this summary of experimental features of the ATC study, it is also important to note the alignment of this study with our survey’s motivating example. The simulator interface is complex and with its multiple regions, the task is dynamic and time-constrained, and subjects had to continually transition their gaze from AOI to AOI.

Now that we have established the linkage between the ATC study and our verification criteria, we can turn our attention to the impact this particular experimental design had on subject performance and eye movement patterns.

Performance data for the ATC task points to significant differences between the two UI conditions: text-based vs. color-coded UI. Subjects accrued significantly more penalty points under the text-based UI and this pattern was exacerbated as the level of task difficulty increased. Surprisingly, the color-coded UI condition lead to user performance of nearly 100% at all levels of difficulty.

Eye tracking data also exhibited significant differences under the two UI treatments. Figures 5a presents a heat map representation for PFT under the text-based UI condition; Figure 5b presents PFT under the color-coded UI. Clearly, under the color-coded UI treatment, there was a significant shift in PFT away from the message panels towards the central region of the UI. Figure 5c captures this result: an approximate 15% shift in PFT. While these results represent averages over the duration of the experiment, it was also observed that changes in the level of difficulty brought about a stepwise decrease in PFT (60%, 57.7%, and 54.8%) away from the aircraft panel (left side of UI) and towards the message panels (right side of UI) under the text-based UI condition. Under the color-coded condition there was a near constant rate of 72% PFT on the aircraft panel throughout all levels of task difficulty.

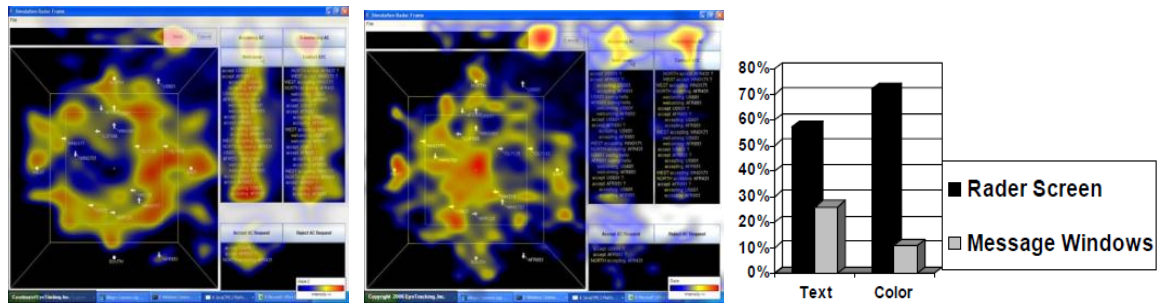


Figure 5 a) Text UI b) Color UI c) PFT results (Bartels and Marshall 2006)

Besides PFT results, the ATC study also discussed findings related to the frequency of transitions between the aircraft panel and the message panels of the simulator. Interestingly, significant differences in transitions per second were observed between the two UI treatments, with subjects having significantly higher rates of transition under the text UI condition. In addition, under both UI treatments, increases in the level of demand lead to a reduction in the rate of transitions. Both of these findings seem to support our belief that the combination of foveal and speed stress are needed in order to maintain the coupling of eye-movement patterns and performance measures.

Combining the results for performance and eye-tracking, we can begin to understand the effectiveness of eye-tracking in providing deeper insights into user behaviors while performing tasks (i.e., the ATC study being an excellent example of what our survey has referred to as the HET perspective). The eye-tracking component of the ATC study identified the attentional dilemma that subjects faced when performing the task under the text-based UI treatment. Namely, subjects needed to both maintain a transition rate from message panel to aircraft region in order to keep pace with changing conditions while also extracting information only available in the text messages. Eye-tracking results thus pointed to the UI feature that was most significant

in improving user performance—color-coding of the aircraft. Because the ATC study gave both eye-movement statistics and task-performance results for each of the UI conditions, it enabled us to see an EPC. Specifically, we were able to see that for this user-interface, there may be a relationship between the transition-frequency and PFT and users' task-performance.

One disappointing aspect of this paper was that the users' performance on the color-coded UI was a little too good. Specifically, users had near perfect performance when they were in the color-coded UI condition. This is somewhat undesirable from the perspective of searching for EPC because it means that there is an overly narrow range of task-performance results in the color-coded UI condition. This can make it more difficult to see statistical relationships between eye-movement metrics and users' task-performance results.

Concluding our analysis of the ATC study, we reiterate some of our observations. First, from the perspective of eye-tracking experimentation, an important take-away message might be that great care should be employed when manipulating the user interface. In this study, a seemingly minor alteration to the UI (color coding the aircraft) led to the significant changes in PFT. While it is commonly understood in HCI experimentation that changes in a user interface can bring about significant changes in performance, what this result indicates is the extreme sensitivity of eye movements to relatively small changes in the user interface.

Second, we find fault in the near perfect task-performance of users under the color UI condition. It is undesirable in an HCI experiment for the task that users perform to be too easy – interesting patterns in the final results may not be apparent under such conditions. When designing HCI experiments, effort must be expended in preliminary studies in order to avoid

these types of unexpected results. As a result of this oversight in the ATC study, the opportunity to compare a range of subject performance data against eye movement measures under the color UI condition was lost.

7.2 IDE Study

In the next paper, the IDE (integrated development environment) study, data from a previous eye-tracking experiment was reanalyzed using a unique approach, in which the data collected for the total duration of the experiment was segmented into smaller time units. Then, the authors compared eye movement behaviors during these time segments. The primary goal of the IDE study was the identification of differences in eye movement behaviors between expert and novice computer programmers using an “integrated development environment” (IDE), which is a piece of software that facilitates the task of computer programming. The authors looked for eye movement patterns from their original experiment that were associated with better performance. However, an important secondary goal of the authors was the demonstration of the technique of temporally segmenting eye-tracking data from an HCI experiment. By increasing the granularity of analysis, the authors hoped to demonstrate a general technique for handling the large datasets typically collected during eye-tracking experiments. As we noted in the introductory sections of our survey, the long duration of most eye-tracking experiments typically leads to much larger data sets than those found in traditional HCI experiments, and the lack of standardized techniques for organizing and analyzing such large datasets has been cited as a

serious impediment to the wider adoption of eye-tracking techniques within HCI research. The IDE study addresses this issue by segmenting the total experimental trial into equal time periods and using the segments as the unit of comparison for eye movement metrics.

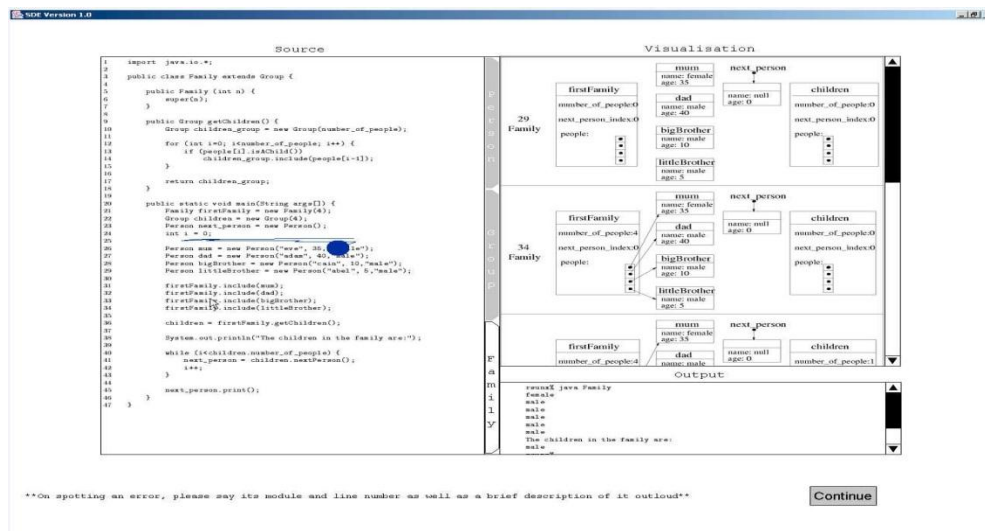


Figure 6 User Interface for IDE paper (Bednarik and Tukiainen 2008).

In the IDE experiment, fourteen (14) programmers, categorized as either expert (n=8) or novice (n=6) based upon months of programming experience, were selected. Subjects performed a software debugging task with the aid of a multi-paneled integrated development environment (IDE); IDEs are used as a tool for managing the complexity of working with the large number of variables, code modules, and their interactions found in a software development project. The IDE contained multiple representations of the program (see Figure 6) including the code (left panel), a visualization of the current state of the program (upper right panel) as well as any program output (lower right panel).

After performing a familiarization task and reading a script pertaining to the correct execution of the program, subjects were given ten (10) minutes to run each of three (3) programs. During this time, subjects searched for a maximum of four (4) logical errors intentionally included in the computer programs by the researchers. During the full ten minutes, subjects' eye movements were recorded. Task performance was evaluated as the number of programming errors correctly identified by the subjects. AOIs were defined over the three panels of the IDE and eye-tracking data was compiled for proportional fixation time (PFT) per each of the three AOIs as well as the total number of transitions and the number of transitions-per-minute between all possible two-way combinations of AOIs (e.g., Code AOI to Visualization AOI (or back)).

This IDE study possesses several of the characteristics of an EPC verification experiment (as defined earlier in this survey). First of all, the researchers followed good experiment design practices: clear subject selection criteria, experimental scripts, and the use of practice trials. Even more significant is the use of a time-constrained task and a UI that required subjects to both closely analyze UI regions for content while simultaneously monitoring the remaining UI regions for asynchronous changes. This is precisely the type of experimental design that we argue is important in creating the foveal and speed stress necessary for the detection of statistical relationships between task-performance and eye movements.

In regard to the task-performance results, the IDE study confirmed that members of the expert group did outperform novices in the debugging task. While this result in itself is not surprising, it does indicate that the authors' criteria used to select subjects was successful in capturing real differences in the abilities of the subjects. This inclusion of both novice and

experts in the experiment provided a healthy range of task-performance outcomes; having a wide variety of levels of task-performance success is desirable in an EPC verification experiment because it may facilitate the search for correlations between eye movements and task-performance.

The most significant eye-tracking research methodology employed in the IDE study was the authors' decision to segment the eye-tracking data into time intervals. As we have noted previously in this survey, a key challenge faced by researchers who want to use eye-tracking data in their experiments is the task of handling the large data set of human behavioral data from the high-frequency eye-measurements of the eye-tracking equipment. The benefits of using a temporal segmentation is clearly demonstrated by comparing the original data of the IDE experiment (Figure 7: Column 2 = Novice PFT; Column 4 = Expert PFT) with the segmented presentation (Figure 8). In Figure 8, we have a finer granularity of analysis, and this reveals important differences in how novices and experts allocate visual attention. Such trends are obscured if only look at the aggregate data (Figure 7).

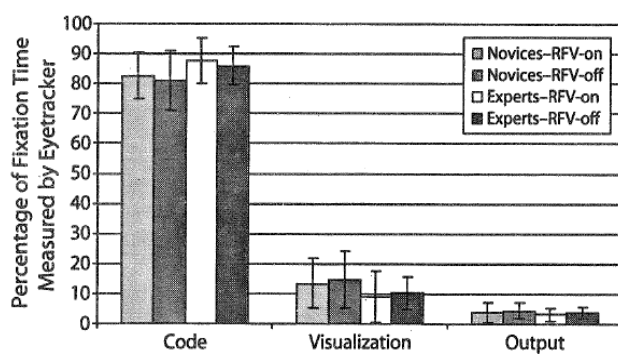


Figure 7 Original aggregated data for IDE study (Bednarik and Tukiainen 2008).

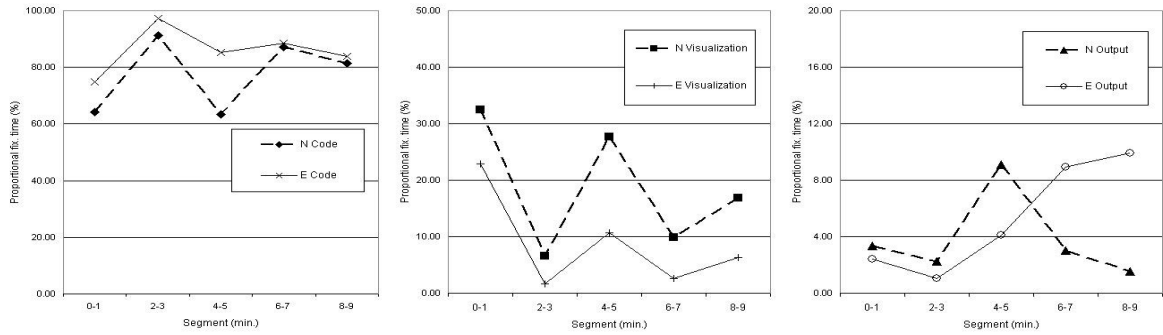


Figure 8 Segmented presentation of PFT (Bednarik and Tukiainen 2008).

In particular, the segmented PFT results reveal that 1) experts have more gradual shifts in their allocation of visual attention across the available AOIs while novices exhibit larger fluctuations; 2) experts have higher PFT on the code view of the program throughout all segments of the experiment; 3) experts display a sharp increase in PFT on the output view during the last time-segment of the experiment while novices are focusing attention on the visualization view. And all three of these eye-movement patterns of experts are associated with superior task performance. Thus, this paper has revealed an eye-tracking/task-performance connection (EPC). If a computer could automatically identify when a computer programmer was being successful or struggling during a debugging task (from analyzing eye-tracking data alone), then perhaps a computer could automatically modify some aspect of the system or provide some additional support. This, identifying this eye-movement / task-performance connection is a valuable finding.

In fact, the observation that expert programmers tended to rely upon the code view of the program significantly more than novices has also been observed in another study included in

our survey. In the LINE study (Section 7.4), a similar pattern of eye-movement was identified and can form the basis of a technique for identifying better task-performers. In the LINE study, which also employed a programming task, it was observed that subjects with higher performance scores exhibited a top-to-bottom reading style at the outset of the debugging task prior to searching the program for errors. As we will see in the discussion of the LINE study, this pattern was then used to construct a visual representation of the eye-tracking data which was associated with superior performance. Because conclusions based on eye-tracking results are typically difficult to generalize across experiments it is encouraging to see two well-designed eye-tracking experiments seeming to confirm each other's findings.

The IDE study's results also reveal a pattern in the users' eye transitions between different AOIs. Not only do experts display significant increases in the overall rate of transitions during the final segments of the experiment, but that superior performers also tended to make rapid transitions between the code and output views. Interestingly, a similar pattern of rapid increase in the rate of transitions associated with better performance was also observed in another study included in our survey (Section 7.3—PILOT study). In the PILOT study, subjects performing simulated aircraft landings displayed sharp increases in the total number of transitions. Thus, a computer, which is monitoring a user's eye transition rate, might be able to automatically identify if a user is tending towards poor task-performance and trigger corrective responses on the UI.

In summary, the IDE study has met the majority of the selection criteria for papers in this survey (Section 6) as well as presenting results which correlate standard eye-tracking measures

with superior performance—or EPC. This study is a superior example of HCI research methodology; it includes: appropriate experimental design, good subject selection, and consistent experimental procedures. In terms of eye-tracking methodology, subjects are exposed to identical visual stimuli, and they performed a challenging and intentional task, while appropriate eye-tracking metrics were gathered. In addition, the UI in the IDE study also has similarities to the motivating example of Section 5; the IDE UI presents dynamically changing content across multiple regions, which requires the subject to continuously make visual transitions from region to region. Thus, the IDE paper has been a valuable piece of HCI research for us to consider in our search for how to best design experiments that can reveal connections between users' eye-movements and their performance on tasks.

7.3 PILOT Study

The original intent of the authors of the PILOT study was to identify eye movement patterns associated with superior performance in landing an airplane; the authors argued that such information could be incorporated into pilot training programs. We will see that, like the ATC and IDE studies, the PILOT study also presents an intriguing relationship between superior task performance and eye transition behavior. Locating such findings is important for our survey in that they both raise the possibility of cross-study comparisons, cited as a significant issue for eye-tracking research, as well as indicating potential areas for future investigation when conducting EPC verification studies.

In the PILOT study experts ($n=7$) and novices ($n=10$) were recruited from two well-defined pools of subjects categorized into groups based upon the following criteria: experts consisted of

fully certified U.S. Air Force pilots with an average actual experience level of 1,980 real flight hours, and the novice group consisted of U.S. Air Force Academy cadets with an average of only 46 virtual flight hours. Each group performed fifteen (15) trials of a simulated airplane landing—three for familiarization and twelve with eye-tracking data recorded. The flight simulator UI (Figure 9) was composed of multiple panels each updated dynamically as the trial progressed with either animations of the exterior view (runway) or data in the interior views (control panel) mimicking a typical airplane cockpit environment.

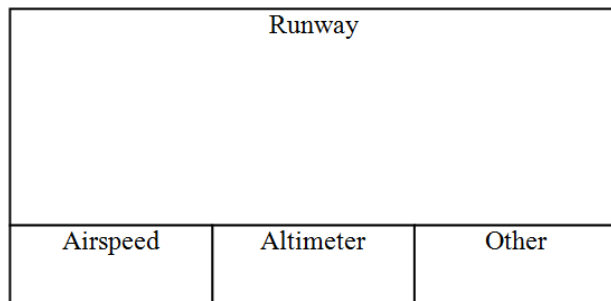


Figure 9 Landing Simulator UI for PILOT study (Kasarskis 2001).

AOIs were defined over each of the four UI regions and the following eye-tracking metrics were recorded: 1) total number of fixations per trial; 2) fixations per individual AOI; 3) transition rate; 4) scan path. Performance was measured on a continuous scale and was calculated as a function of the distance from the optimal landing point on the airport runway represented by the cross-hairs in Figure 10.

From an HCI experimental perspective, the authors of the PILOT study did a good job of handling of how the novice and expert groups were constructed. By drawing from legitimate

populations of subjects (actual pilots) and not relying on a “convenient sample” (i.e., “students in our academic department”) the findings of the PILOT study will have greater credibility when it comes to connecting the performance and eye-tracking results.

The PILOT study has several characteristics of an EPC verification experiment, which we introduced in Section 6. The visual stimuli presented to each subject were well-controlled with each subject seeing the same interface and external scene. Also, the task included a continuous performance success scale, and the subjects in the study had a variety of levels of success. Thus, the experiment produced a healthy range of numerical success scores of the participants – thereby facilitating the search for statistical relationships between eye movements and task-performance. Another desirable property of the PILOT study is that, like our motivating example in Section 5, it included a dynamic task (complex, time-constrained) and a visual dispersion on the UI (subjects must transition their visual attention in order to complete the task). We believe that an experiment possessing both of these properties is better able to generating the foveal and speed stresses necessary to induce tunneling in the subjects. As we have argued earlier, the tunneling users’ experience when they feel overwhelmed or overloaded during a task may lead them to move their eyes less – because they tend to have less awareness of regions of the visual field outside of their current focus. Thus, this may lead to a relationship between something an eye-tracker can detect and some internal mental state or frustration-level of the user. Combining all of these factors together, we argue that the PILOT study has many of the properties of a good EPC verification experiment; thus, we were hopeful when reading this paper that we would be able to find some evidence of a connection between task-performance and eye-tracking data.

In terms of the actual performance of subjects, Figure 10 provides a visual display of the differences between expert and novice pilots. The landing patterns reveal that fully qualified pilots perform better landings than novice pilots; however, what is significant to our survey is that the authors' care spent on the subject selection process not only translated into measurable differences in performance outcomes, but also provided a wide range of outcomes which will provide better correlations with the eye-tracking measures.

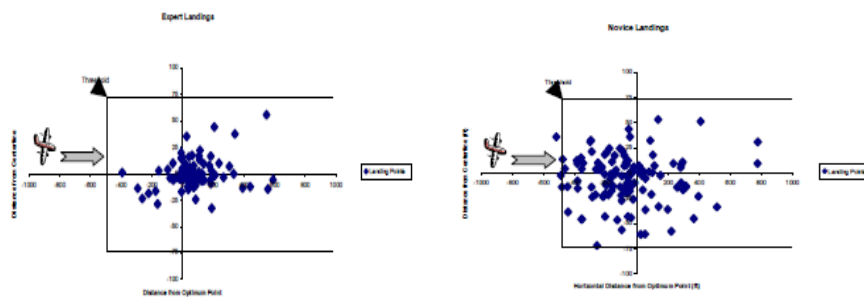


Figure 10 Pilot Landings a) Experts (on left) b) Novices (on right)

The first eye-tracking result we present reveals a relationship between eye-movements and task-performance. Figure 11 shows typical scan paths for expert and novice pilots. The first observation we make about the scan paths is that in both cases, novice and expert, there is a noticeable organization in the patterns. (Users are not wandering their eyes all over the user-interface with no recognizable pattern.) This is a result of the authors' providing the necessary stress on the visual attention system through a combination of dispersion in the UI and the dynamic nature of the task. In the absence of this foveal stress, we would have expected to see greater randomness in the pattern and accumulations of fixations unrelated to task performance. The second observation we make is that there are distinguishable differences in the scan paths associated with the level of expertise. As Figure 11 demonstrates, experts tended to have better

organized scan paths displaying greater economy in their use of the UI while the novices exhibited noisier eye movement patterns. Because this PILOT study was able to produce data that consisted of both (1) organized eye-movements during the completion of a challenging task and (2) a good variation in task-performance scores from different users, we were successful in searching for relationships between these two variables (eye movement, task-performance). It is precisely this zone of eye movement pattern and performance results that an EPC verification experiment must generate if it is to be capable of detecting connections between task-performance and eye-tracking measures.



Figure 11 a) Expert (Left) b) Novice (Right) Sample Scan Paths (Kasarskis 2001).

Further characterization of eye movement patterns of the two groups is provided by an analysis of the average dwell time (how long did a series of fixations accumulate in an AOI before visual attention moved to another AOI) and the count of total fixation (Figure 12). What this result indicates is that experts have developed greater levels of automaticity in their scanning behaviors and are able to extract information at higher rates than is exhibited by the novices. Stating this finding another way, we can say that subjects with higher levels of expertise also

have higher transition rates. The question then becomes: Is there a connection between higher rates of transitions and performance?

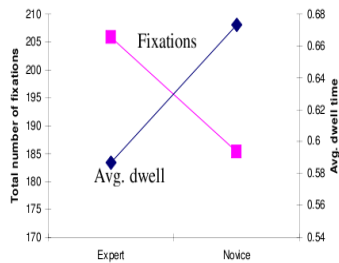


Figure 12 Total Fixations/Average Dwell Time Experts and Novices (Kasarskis, Stehwien et al. 2001)

Figure 13 provides an answer to this question and indicates that, within each of the groups, better performance was indeed associated with the higher transition rates.

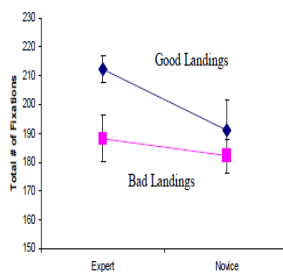


Figure 13 Fixations and Performance (Kasarskis, Stehwien et al. 2001)

And this is the intriguing finding that we mentioned in the introduction to the PILOT study discussion—higher transition rates associated with better performance have been observed in the ATC study, in the IDE study, and now in the PILOT study. Given that studies on diverse user-interfaces (with diverse original intents of the authors) were able to reveal such a pattern, it will be interesting to see if future EPC verification experiments on other user-interfaces are able to identify similar trends.

In summary, important characteristics of the PILOT study include the following: 1) well-designed experimental procedures including clear subject selection criteria, 2) precisely specified performance measures, 3) a task/UI combination capable of generating the necessary speed and foveal stresses needed in order to generate a range of user outcomes. All of these features contributed to the PILOT study having properties of an EPC verification experiment. Further, looking at the data from the PILOT study, we were able to see connections between eye-movement patterns and task-performance results. Thus, this study is another example of an experiment that has been successful in revealing a statistical connection between patterns of eye-movement and task-performance – despite this not being an explicit intention of the authors when they began their study.

7.4 LINE Study

Having completed three analyses of eye-tracking studies, the validity of Jacob and Karn's critique should be clear—eye-tracking research lacks effective methods for handling the large amounts of human behavioral data collected during experiments. Therefore, studies which have developed unique methods of data analysis which disaggregate or transform eye-tracking data into more useful formats have been praised in this survey. As we discussed in the IDE study, one possibility for dealing with this data issue is to subdivide the data temporally and then aggregate standard eye-tracking metrics across time segments. The benefit of using the temporal divisions was that it uncovered eye movement patterns that would have otherwise been obscured. In the current discussion of the LINE study, we present the findings of an eye-tracking study which also attempts to address the data disaggregation issue.

In the LINE study, a novel eye-tracking metric is constructed by discarding data (specifically, the authors retain only the vertical dimension of the eye-tracking data) with the goal of transforming the complex scan path metric (see Figure 11 in Section 7.3) into a simpler format. Importantly, not only did this technique provide an effective method for characterizing eye movement behavior, but specific patterns emerged which were found to be associated with better task performance.

In terms of experimental design, the LINE study consisted of five (5) volunteer subjects selected from a computer science program based upon their familiarity with both programming languages and techniques of formal code review. All subjects were graduate students and the group had a mean of 3.5 years of programming experience. The task employed in the LINE study was similar to that used in the IDE study except in the LINE study each program contained only one error. The user interface consisted of a single paneled UI which only allowed subjects to read the program. In the LINE study, there was no means to execute the program, and the UI did not provide any visualization or program output.

Experimental trials began with the reading of a script explaining the intended function of the computer program. During this introduction subjects were told that each program contained a single error and that they had five minutes in which to read and locate the error. Each subject analyzed six programs containing between 10 and 20 lines of code. Eye-tracking data was recorded for all 30 trials and task performance was measured as the clock time until the error was detected.

Earlier in this survey, we discussed in the importance of designing experiments that produce sufficient speed/foveal load such that the user is forced to make specific types of eye movements at a rigorous speed – if you are seeking relationships between eye movement and task performance. In the LINE study, the simplicity of the UI (a single-panel UI) and the use of a less complicated user task (identification of a single programming error) might not seem capable of providing the necessary foveal stress in order to establish a correlation between eye-tracking data and user performance. However, as we will argue, on closer inspection, both of these features led to a tighter coupling of eye movement measures and performance data which point to important lessons in the design of verification experiments.

The apparent lack of complexity in the interface was counterbalanced by a novel method of defining areas-of-interest. In the LINE study, each line of text in the program was defined as a separate AOI with eye-tracking data collected per line. In the next step of this approach, the authors modified the standard scan path data so that it only included the vertical displacement (line-to-line motion) of the eye movement while discarding all lateral movement. In this discussion we will refer to this simplified scan path as the DISCRETE-VERTICAL path (because they discard the horizontal eye movement information and they discretize the vertical information into individual strips that correspond to each line of the computer program). The effect of this simplification was to transform what is typically a very complex eye-tracking metric into a visual representation more closely linked with the specific task being performed (compare Figure 11 in Section 7.3 with Figure 14 below).

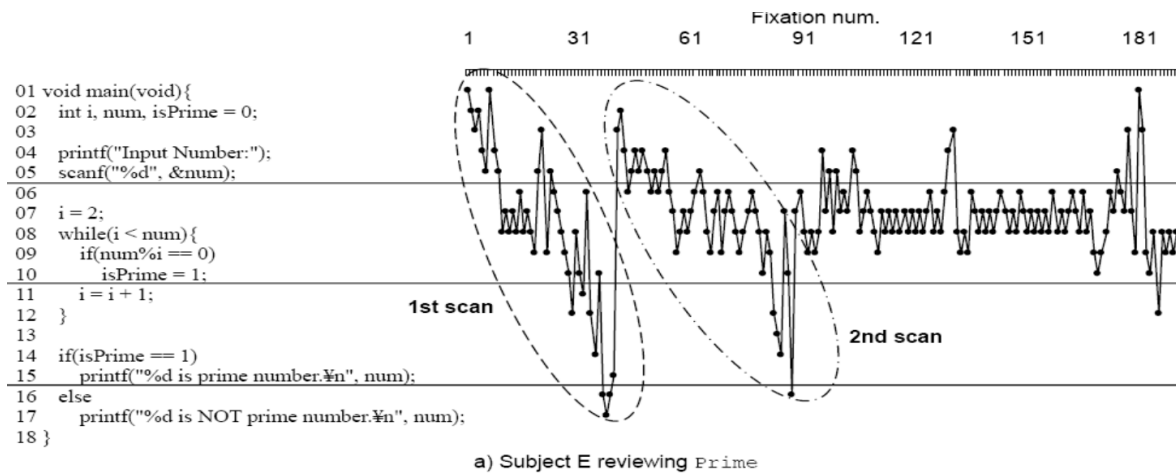


Figure 14 Example of a modified scan path in the LINE study (Uwano, Nakamura et al. 2006).

The significance of the DISCRETE-VERTICAL technique is that it demonstrates another method for handling the large data sets produced in eye-tracking experiments. By discarding large amounts of unnecessary complexity in the data the DISCRETE-VERTICAL format revealed eye movement patterns that would otherwise not have been observable in the raw scan path data. Specifically, it was observed that subjects with a DISCRETE-VERTICAL scan path that covered 70% of the lines in the program during the first 30% of the time of the experimental trial were also more likely to discover the programming error more quickly.

The second feature of the LINE study experiment that we wish to draw attention to is the use of a task which only consists of one unit of work. In contrast to both the ATC and IDE studies, where subjects repeatedly performed a subtask (i.e., handle multiple aircraft, locate multiple programming errors), in the LINE study, eye-tracking data was collected for a single iteration of the task. We argue that if the goal is to design experiments that attempt to correlate eye-tracking data with a performance measure then it is important to guarantee that the collected

data is aligned with the starting and finishing boundaries of a single task performance. Specifically, if we are collecting performance scores for individual sub-tasks during an experiment, it is desirable for us to know which time slice of the eye-tracking data corresponded to the user's work on that sub-task. In the experimental design of the LINE study, it is easy to see this link between portions of the eye-tracking data and portions of the tasks (and the user's success) because there is only one "error" to be found in each computer program shown to participants. In contrast, in the IDE study, the user is searching for multiple "errors" in a computer program, and so when later analyzing the eye-tracking data, it would not be clear which part of the eye-movements corresponded to the user's efforts to find each of the "errors." As a consequence, because the user's efforts to find the "errors" in the program may be more interleaved, the correlation between eye-tracking data and the user's success at finding individual "errors" may be more difficult to uncover.

In criticism of the LINE study, there might have been better, more subtle approaches to filtering the eye-tracking data than merely discarding the lateral movement data for each line of the program. In the programming task, it might be even more significant to examine where on the line subjects looked – as opposed to merely examining which line they looked at. The assumption made by the authors of the LINE study is that information critical to the identification of errors in a computer program is uniformly distributed across each line of the program – when, in fact, it is more likely that certain areas of a line in a computer program have a higher probability of containing an error (e.g., data types in a method header).

In summary, the LINE study makes significant contributions to our understanding of what it takes to conduct a successful verification experiment model (an experiment which allows us to search for connections between eye-movement patterns and users' task-success). By demonstrating a novel technique of manipulating the eye-tracking data, the LINE study addresses the fundamental issue of how to handle the large data sets generated from eye-tracking experiments. Besides providing a compact description of subject performance, the DISCRETE-VERTICAL format of the scan path also allows for better correlations between performance and eye-tracking metrics. Given the theme of this survey, it is exciting that the LINE study identified an explicit connection between eye-tracking movements and task-performance – i.e., the relationship between task-success in this experiment and the user performing a 70% DISCRETE-VERTICAL scan path during the first 30% of the experiment time.

7.5 NEWS Study

In the final paper of our survey, we explore an eye-tracking study which had the original research goal of measuring the impact on task performance brought about by changes in the user interface and to attempt to relate these results to eye-tracking measures. As an eye-tracking study that explicitly focused on the relationships between user interface, task performance, and eye-tracking measures, the NEWS study was clearly of interest to this survey. Unfortunately, the NEWS study had deficiencies in its experimental design, which may explain the weak connections observed between the eye-tracking measures and user performance. These defects in the experimental design included both the simultaneous use of multiple information formats

(e.g., audio and video) – which inadvertently reduced foveal stress – as well as poor control of the presentation of visual stimuli – which interfered with the collection of eye-tracking data. These shortcomings in the NEWS study were especially disappointing because its interface/task combination had strong similarities to our motivating example (Section 5) – so, if the experimental design had been better, we would have been very interested in the results. Nevertheless, we will use the negative aspects of this study as examples of what to avoid in any future verification experiments that we might perform.

In the NEWS study the experimental design consisted of a sample of undergraduate students (males= 23, females=13) drawn from the campus population with a mean age of 24.3 years. The task consisted of watching three television news stories ($\approx$ 2 minutes in length) taken from foreign media sources (in order to limit the possibility of prior viewing by the subjects). User performance was measured by post-testing subject’s ability to recall factual information.

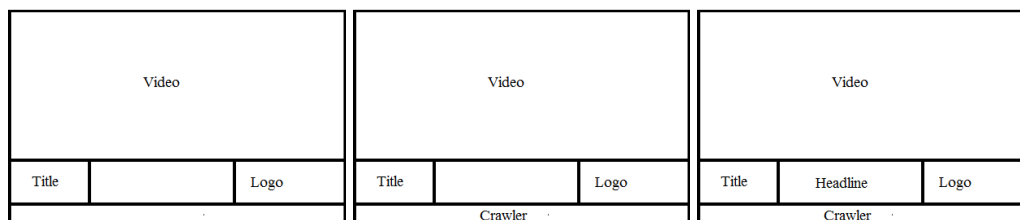


Figure 15 Three versions of the NEWS study UI a) Base version b) With Crawler c) Both Crawler and Headline (Josephson 2006).

Three versions of a user interface were employed in the NEWS study with adding an additional visual element (see Figure 15). A base version consisted of only the main video region, while a second version consisted of the base version plus a text crawler that presented information unrelated to the video. A final version added both the crawler and a headline region

to the screen. In contrast to the content in the crawler, the purpose of headline region of the screen was to reinforce specific facts contained in the main video. All of the versions of the UI contained a generic title and globe (station logo) elements.

The intent of this experimental design was to test the authors' hypothesis that highlighting factual content in the headline area would reinforce learning of information from the main story being presented – while the text crawler, by providing a distraction, would interfere with learning and thereby lower task performance. As would be expected, areas-of-interest were defined over UI regions which were considered to be important to task performance (main, headline, and crawler), and visual attention was measured in terms of proportional fixation time (PFT) per AOI and scan path data.

Several aspects of the way the experimental study were conducted would suggest that the NEWS study would be a likely to be successful at identifying a connection between eye-tracking data and task-performance. Specifically, the authors conducted several aspects of the experiment well: good subject selection, the use of experimental scripts, and clear performance measures – which points to a carefully thought out experimental design. More importantly, for the purposes of our verification experiment design, the user interface has a degree of visual dispersion in line with the other eye-tracking studies we have reviewed: the user needs to look at different regions of the screen when doing the task. Further, the task appears to be of a dynamic nature which will require subjects to keep pace with the presentation of the material. As we have argued before, both of these experimental factors are needed in order to provide the foveal and speed stresses necessary to keep the subjects visual attention focused on task performance.

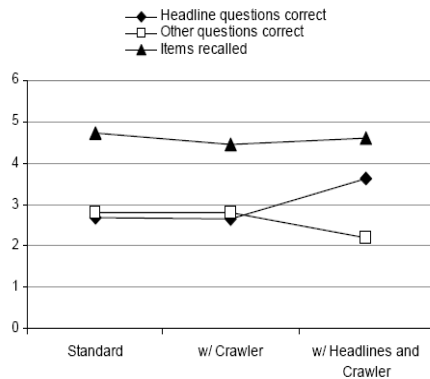


Figure 16 Performance results for NEWS study (Josephson and Holmes 2006).

However, when we begin to analyze the performance and eye-tracking results of the NEWS study we begin to see where problems arise in our ability to correlate performance outcomes with eye-tracking measures. As we have argued, in the design of verification experiments it is necessary to generate healthy range in performance between groups (i.e., a good “spread” in the results) in order to allow for the detection of correlations between performance and eye-tracking measures. When all of the performances results are similar, then it is harder to find these correlations. The performance results of the NEWS study (Figure 16) do not provide clear separation between the different user interfaces. Looking at the upper line in Figure 16, we see that there are only small differences in the ability of subjects to recall information under the various interface conditions. These results of the NEWS study can be contrasted with those in the ATC study under the two user interface conditions (text-based vs. color-coded) used in that study – in the ATC study wider separation in performance scores across conditions was observed.

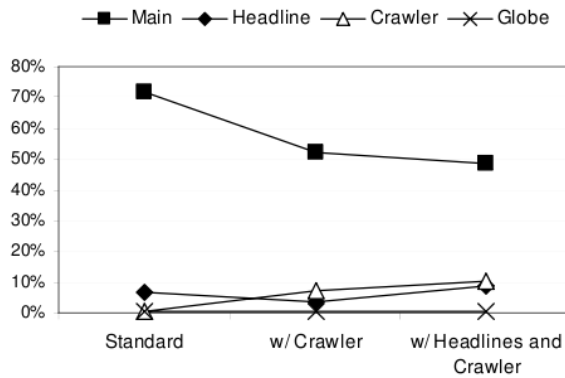


Figure 17 PFT results for NEWS study (Josephson and Holmes 2006).

Moving the discussion to the eye-tracking results (Figure 17), we see that large shifts in proportional fixation time occurred as a result of the inclusion of the crawler (upper line Figure 17). From our perspective of designing verification experiments, it would have been great if this dramatic change in eye-behavior had been accompanied by a corresponding change in the task performance scores for that user-interface condition. If that had occurred, then we would have found an eye-movement behavior with a strong link to task-performance. Unfortunately, it appears that there is little or no connection between performance and PFT in this regard.

In providing an explanation for why the NEWS study did not identify any eye movement behavior related to task performance, we believe that a likely explanation is that the user interface contained multiple modes of presenting the information content. In particular, the NEWS study (unlike all other eye-tracking studies in our survey) also contains audio – in addition to the visual stimuli. For this reason it is likely that subjects had greater freedom in allocating visual attention during experimental trials because they could gather the same information by just listening to the audio track. In the extreme, it might be possible for a subject working with such an interface to literally close their eyes and just listen. As a result of this lapse

in the experimental design, it would seem very unlikely that a close connection between eye-tracking measures and user performance could have been identified.

We have discussed previously that sometimes large changes in eye-tracking measures are the result of small changes in the user interface. One aspect of this NEWS study further supports this point: there is the large shift in PFT after the addition of the crawler to the UI. The addition of this small visual element led to a significant change in eye behavior. This sensitivity of eye-movement behavior to small UI changes is part of why it is challenging to design user-interfaces experiments that include an eye-tracking component. Eye movement patterns can be very sensitive to even small changes in the visual stimuli presented to subjects, and for this reason it is important that researchers are aware of the possibility of producing large and unexpected fluctuations in eye-tracking data with the introduction of what appear to be innocuous visual elements.

As a final point, we have stressed earlier in our survey that controlling for the presentation of visual stimuli is an important aspect of designing experiments that include eye-trackers. One form of control is the order in which subjects view different user interfaces throughout an experiment. For example, consider the scenario in which subjects first view an interface that includes the headline region, and then later in the study, the same subject views an interface without the headline region. In this situation, it would not be surprising if the subject had developed an expectation (a learning response) that the headline region should be present. In such a scenario, the eye-tracker would record fixation data for a non-existent AOI – interestingly, Figure 17 (columns 1 and 2) indicates that this was occurring in the NEWS study. Another

aspect of the experimental design of the NEWS study that indicates a lack of control in the visual stimuli is in the fact that they experiment included a wide diversity of video news content in the different user-interface conditions. Some types of visual stimuli (e.g., faces, horizon lines) contained in the videos may have attracted user's eye-gaze more than others, and this could have affected the data collected in this study.

In summary, the NEWS study is an important demonstration of some of the difficulties faced in designing eye-tracking experiments in general, which in turn also had a negative impact on correlating user performance with eye-tracking data. In particular, we observed that controlling for the presentation of visual stimuli is perhaps a more subtle challenge facing researchers than is widely recognized. Of all the papers in our survey, the NEWS study argues for the need of carefully developed guidelines for the design of verification experiments.

8 Summary of Findings

Having concluded the paper critique section of the survey, we would now like to summarize our findings. Before doing so we would like to remind the reader that while we have discussed the papers from both an experimental HCI (HET) perspective and also from an Eye-tracking/Performance Connection (EPC) perspective, the primary goal of our survey was identifying the necessary experimental conditions which we argued would lead to connections between eye-tracking measures and task performance. In this section we will provide a set of guidelines or take away points that highlight what we consider to be the most important points that would aid in the design of any such future verification experiments.

The first finding clearly demonstrated in our survey is the sensitivity of eye movements to small changes in visual stimuli. Unlike other human factors such as physical response time and even other biometric measures such as heart and blink rate which under normal conditions respond to stimuli in stable and predictable limits, eye movements are extremely sensitive to changes in experimental conditions. In the experiments we have reviewed we have observed 20% fluctuations in PFT (ATC, NEWS) generated by only small modifications to the UI. Additionally, patterns of eye movement behavior seem to be quickly established after only short exposures to new visual stimuli and these patterns continue even when the expected elements are removed from the stimuli (NEWS). These studies indicate that in order to conduct successful verification experiments additional care must be taken in handling of visual stimuli and how it is presented to subjects. This observation also points to areas that verification experimental design might differ from HCI experiments. In particular, while between-subjects designs might be appropriate in experiments which do not change the visual stimuli which subjects are exposed to, in eye-tracking experiments in which do employ multiple UIs it would be more appropriate to use a between-group design thus avoiding the negative impacts on eye-tracking results as seen in the NEWS study.

A second lesson learned addresses the issue raised in Section 4 of our survey relating to the size and complexity of data sets generated during eye-tracking experiments. Namely, unlike more traditional HCI experiments which measure human response over relatively short time intervals, the eye-tracking experiments discussed in our survey averaged approximately 7 minutes. In addition, the tasks were of a more complex nature and could reasonably be seen as consisting of stages. Given these factors it is probable that aggregating eye-tracking data over the

total length of the experiment will obscure any shift in eye movement patterns which occur during the experiment. In our survey we identified two papers which provided innovative approaches in dealing with the difficulties inherent in analyzing large eye-tracking data sets (IDE and LINE). In the IDE study the data were segmented into fixed units of time which was fine-grained enough to isolate the timing of shifts in eye movement behavior between experts and novices. In contrast, the LINE study segmented the data spatially which allowed for the identification of a pattern related to user success. The take-home lesson of these examples is not that there is a single best technique in segmenting the data sets, but that some ingenuity on the part of the researcher will be necessary to uncover the eye movement patterns concealed in the large data sets.

A third lesson was the importance of properly designing the task. Our survey has indicated the importance of controlling for both task type and task difficulty. In the NEWS study we have an example of what results from the use of a passive task type with low speed-stress and foveal load. Under such conditions subjects are free to allocate their visual attention across the UI with little connection to the task they are performing. As a result, eye-tracking measures are poor indicators of subject performance. On the other hand, the ATC and IDE studies provide examples of active tasks which contain both high speed stress and foveal loads. In this case, it was observed that changes in performance could be correlated with changes in an eye-tracking measure (transition frequency).

A fourth lesson our survey is the importance of UI complexity as modeled by our Motivating example. Such interfaces force the subject to continuously transition visual attention

from one area of the UI to another and in doing so provide a level of speed and foveal stress which in combination with a demanding task does not allow subjects to vary attention to non-task specific UI elements. Just as was observed in the importance of task definition, this too contributes to cleaner eye-tracking data.

In addition to these major points we have also seen examples of experimental practices which we believe have contributed to the quality of the eye-tracking experiments presented in our survey. For example, if the eye movement patterns of groups with varying expertise are under study then it is highly important that clear measure of this difference in expertise be employed such as was used in the PILOT study.

However, the main take-away message that our survey makes is that in eye-tracking (due to the great sensitivity of eye movements to change) there is a complex interplay between all the experimental factors which differentiates eye-tracking research from other areas of HCI experimentation. Eye-tracking researchers must therefore pay greater attention to the experimental concerns that our survey has presented if they are to move eye-tracking research forward.

9 Conclusions

The results of our survey present a snapshot of the types of applications of eye-tracking in HCI studies related to usability testing over the past decade. In examining these results we have catalogued many of the difficulties that are encountered when attempting to use eye-tracking within the context of HCI experiments. We have also seen that while eye-tracking has been in

use for many decades it is still treated as an adjunct to other experimental techniques and has not developed into a primary technique in the area of usability testing. Our survey has argued that this is due in part to a lack of standardization in eye-tracking experiments which results from a lack of understanding of the limits of this technology. This view is in line with recent surveys of eye-tracking which characterize it as a research tool which is always falling short of its promises in delivering a superior tool for evaluating human performance within HCI experimentation (Jacob and Karn 2003).

However, in contrast to this pessimistic view of eye-tracking, our survey has argued that if key areas of eye-tracking experiments are handled appropriately then it is possible to deliver results which allow for direct connections to be established between eye-tracking and user performance. And as we stated in Section 1 of our survey, the importance of establishing this connection between eye-movements and performance is that it would open possibilities of system automation. Namely, it could allow computer scientists to make use of software that automatically looks for these patterns of eye behavior (the observable data) and then make decisions about the user's internal state or aptitude at a task (the hidden data).

Therefore, the goal of our survey was the marshaling of evidence to support connections between eye-tracking and human performance or what we referred to as the EPC perspective. We believe that our survey has provided a preliminary codification of the most important experimental factors that contribute to this connection and therefore have provided a framework for future attempts at verification experiments.

10 Future Directions

Perhaps one of the greatest challenges to eye-tracking research is how to handle the “noisy” data generated from human eye movement data. As we saw in many of the studies included in our survey a great deal of energy and ingenuity was expended in transforming the raw eye-tracking data into a format that would allow for easier analysis and interpretation. Furthermore, most of the analyses performed in the studies included in our survey were done in a top-down fashion and that was driven by various models of human performance.

However, given the nature of eye-tracking data it has been recognized that bottom-up, data-driven approaches might be more appropriate for handling this human behavioral data. In particular, the recent proceedings of the Eye Tracking Research and Applications Conference (ETRA) have included many examples of machine learning (ML) techniques being applied to eye-tracking data.

One area in particular is the analysis of user interactions with Information Retrieval (IR) systems. Approaches within this area have attempted to use eye movement data as a predictor of the relevance of both textual and image search results (Puolamaki, Salojarvi et al. 2005; Hardoon and Pasupa 2010). Acting on the detected level of relevance, systems can be created which improve the efficiency of the search process by reducing the number of iterations that a user must go through in order to find suitable results. Building on this framework it should be feasible that other user states (e.g., frustration, boredom) could be detected with similar ML approaches.

A second area for future directions is that of the prediction of human error (Ratwani, McCurry et al. 2008). In this approach ML techniques analyze a stream of eye-tracking data and determine the probability of human error occurring. If the probability is above a threshold then UI actions occur which alert the user to the error condition and provide a path way to a correct action.

A third area of future directions for eye-tracking is in the area of applications which address users from special populations such as Deaf users. While some early work has already been done in this area it has lacked the experimental rigor that our survey has featured as being critical to good eye-tracking analysis (Cavender, Bigham et al. 2009).

Future directions of our research might include many of these approaches in the design of new user interfaces within educational learning systems designed to address the needs of special populations. However, prior to this we will need to design and conduct careful experiments which look for the relationships between eye movement patterns and user task performance. It is for this reason that we have analyzed and discussed the papers in this literature survey.

11 References

- *Bartels, M. and S. P. Marshall (2006). Eye tracking insights into cognitive modeling. Proceedings of the 2008 Symposium on Eye Tracking Research & Applications. San Diego, California, ACM.
- *Bednarik, R. and M. Tukiainen (2008). Temporal eye-tracking data: evolution of debugging strategies with multiple representations. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. Savannah, Georgia, ACM: 99-102.
- Cavender, A. C., J. P. Bigham, et al. (2009). ClassInFocus: enabling improved visual attention strategies for deaf and hard of hearing students. Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility. Pittsburgh, Pennsylvania, USA, ACM: 67-74.
- Cooke, L. and E. Cuddihy (2005). Using eye tracking to address limitations in think-aloud protocol. IPCC 2005. Proceedings. International Professional Communication Conference, 2005.
- Cuddihy, E., Z. Guan, et al. (2005). "Protocol Considerations for Using Eye-Tracking in Website Usability Testing."
- Duchowski, A. (2002). "A breadth-first survey of eye-tracking applications." Behavior Research Methods, Instruments, & Computers **34**(4): 455-470.
- Eger, N., L. Ball, et al. (2007). Cueing retrospective verbal reports in usability testing through eye movement replay. HCI 2007, Proceedings of HCI 2007.
- Guan, Z., S. Lee, et al. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. Proceedings of the SIGCHI conference on Human Factors in computing systems. Montr\&\#233;al, Qu\&\#233;bec, Canada, ACM: 1253-1262.
- Hammoud, R. I. and J. B. Mulligan (2008). Introduction to Eye Monitoring. Passive Eye Monitoring. R. I. Hammoud, Springer Berlin Heidelberg: 1-19.
- Hardoon, D. R. and K. Pasupa (2010). Image ranking with implicit feedback from eye movements. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. Austin, Texas, ACM: 291-298.
- Helleberg, J. R. and C. D. Wickens (2003). "Effects of Data-Link Modality and Display Redundancy on Pilot Performance: An Attentional Perspective." The International Journal of Aviation Psychology **13**(3): 189 - 210.
- Hewett, B., Card, Carey, Gasen, Mantei, Perlman, Strong and Verplank (1992). ACM SIGCHI curricula for human-computer interaction, ACM: 162.
- Hornbaek, K. (2006). "Current practice in measuring usability: Challenges to usability studies and research." Int. J. Hum.-Comput. Stud. **64**(2): 79-102.
- Hughes, G. and P. Robinson (2007). Photonote evaluation: aiding students with disabilities in a lecture environment. Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility. Tempe, Arizona, USA, ACM: 99-106.
- ISO (1998). "Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability."

- Jacob, R. and K. Karn (2003). "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises."
- Jacob, R. J. K. (1995). Eye tracking in advanced interface design. Virtual environments and advanced interface design, Oxford University Press, Inc.: 258-288.
- Jacob, R. J. K. and K. S. Karn (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. The Mind's Eye (First Edition). J. Hyönä, R. Radach and H. Deubel. Amsterdam, North-Holland: 573-605.
- *Josephson, S. and M. E. Holmes (2006). Clutter or content?: how on-screen enhancements affect how TV viewers scan and what they learn. Proceedings of the 2008 Symposium on Eye Tracking Research & Applications. San Diego, California, ACM.
- *Kasarskis, P., Stehewien, J., Hickox, J., Aretz, A., Wickens, C. (2001). Comparison of expert and novice scan behaviors during VFR flight. 11th International Symposium on Aviation Psychology.
- Kowler, E. (2006). "Attention and Eye Movements."
- Medina, E., E. Cuddihy, et al. (2008). "USES OF EYE TRACKING TECHNOLOGY IN DESIGN." Human Factors and Ergonomics Society Annual Meeting Proceedings 52: 1574-1578.
- Poole, A. and L. J. Ball, Eds. (2005). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. Encyclopedia of Human-Computer Interaction.
- Puolamaki, K., J. Salojärvi, et al. (2005). Combining eye movements and collaborative filtering for proactive information retrieval. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. Salvador, Brazil, ACM: 146-153.
- Ratwani, R. M., J. M. McCurry, et al. (2008). Predicting postcompletion errors using eye movements. Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. Florence, Italy, ACM: 539-542.
- Rayner, K. (1998). "Eye movements in reading and information processing: 20 years of research." Psychol Bull **124**(3): 372-422.
- Richardson, D. C. and M. J. Spivey (2004). "Eye Tracking: Characteristics and Methods." Encyclopedia of Biomaterials and Biomedical Engineering: 568 - 572.
- Roberts, V. L. and D. I. Fels (2006). "Methods for inclusion: Employing think aloud protocols in software usability studies with individuals who are deaf." Int. J. Hum.-Comput. Stud. **64**(6): 489-501.
- Sibert, J. L., M. Gokturk, et al. (2000). The reading assistant: eye gaze triggered auditory prompting for reading remediation. Proceedings of the 13th annual ACM symposium on User interface software and technology. San Diego, California, United States, ACM: 101-107.
- *Uwano, H., M. Nakamura, et al. (2006). Analyzing individual performance of source code review using reviewers' eye movement. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. San Diego, California, ACM.
- Williams, B. (1985). "Tunnel Vision Induced by a Foveal Load Manipulation." Human Factors: The Journal of the Human Factors and Ergonomics Society **27**: 221-227

