

**Cleaning up systematic error in eye tracking
data by using required fixation locations**

Anthony J. Hornof and Tim Halverson
Department of Computer and Information Science
University of Oregon

Email addresses:
hornof@cs.uoregon.edu
thalvers@cs.uoregon.edu

Mailing address and phone number:
Department of Computer and Information Science
1202 University of Oregon
Eugene, Oregon 97403-1202
(541)346-1372

Submitted to the journal
Behavior Research Methods, Instruments, and Computers
Special Issue on Eye Movement Research
First draft submitted: 3/15/02
Revised draft submitted: 9/3/02
Final version submitted: 9/25/02

Word count: 8,500

This work was supported in part by the Office of Naval Research through grant
N00014-01-10548 to the University of Oregon, Anthony J. Hornof, principal investigator.

ABSTRACT

In the course of running an eye tracking experiment, one computer system or subsystem typically presents the stimuli to the participant and records manual responses, and another collects the eye movement data, with little interaction between the two during the course of the experiment. This article demonstrates how the two systems can interact with each other to facilitate a richer set of experimental designs and applications, and to produce more accurate eye tracking data.

In an eye tracking study, a participant is periodically instructed to look at specific screen locations, or *explicit* required fixation locations, such as to calibrate the eye tracker to the participant. The design of an experimental procedure will also often produce a number of *implicit* required fixation locations (RFLs)--screen locations that the participant must look at within a certain window of time or at a certain moment in order to successfully and correctly accomplish a task, but without explicit instructions to fixate those locations. In these windows of time or at these moments, the disparity between the fixations recorded by the eye tracker and the screen locations corresponding to implicit RFLs can be examined, and the results of the comparison can be used for a variety of purposes. This article shows how the disparity can be used to monitor the deterioration in the accuracy of the eye tracker calibration, and to automatically invoke a recalibration procedure when necessary. This article also demonstrates how the disparity will vary across screen regions and participants, and how each participant's unique *error signature* can be used to reduce the systematic error in the eye movement data collected for that participant.

INTRODUCTION

Eye trackers are devices that record gazepoints. A *gazepoint* is a location towards which the eyes are pointed at a moment in time. Eye trackers typically record a person's gazepoint every 1 to 17 ms. When a series of gazepoints occur near each other in time and location, they are aggregated and assumed to represent a single *fixation*, a brief period of time lasting about 100-400 ms during which the eyes are held reasonably stable and steady at a location. A fixation typically ends abruptly with a rapid ballistic eye movement, or *saccade*, that moves the eyes to a new fixation. (Rosenbaum, 1991)

Problem #1: Eye trackers exhibit variable and systematic error.

In capturing each individual gazepoint, eye trackers exhibit two types of error--*variable* and *systematic*. Both can be seen in the gazepoints shown in Figure 1. Variable error can be seen in the spread (or *dispersion*) of the recorded gazepoints around the actual fixation maintained by the human. Systematic error is the disparity (or *drift*) between the average gazepoint location and the actual fixation maintained by the human. Variable error indicates a lack of *precision*. Systematic error indicates a lack of *accuracy*.

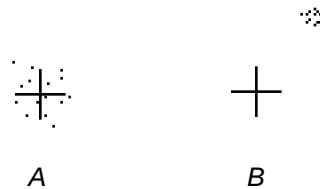


Figure 1. The raw gazepoint data recorded by eye trackers exhibit two types of error--variable and systematic. Hypothetical participants A and B were instructed to fixate a set of crosshairs. The gazepoints for A show no systematic error but large variable error. The gazepoints for B show a large systematic error but small variable error. (Derived from Chapanis, 1949)

Variable error can be reduced by improving the eye tracking machinery (i.e. by using a head-mounted instead of a remote unit) and experimental method (i.e. by using a bite guard). Variable error is also removed, at least from view, by a fixation-detection algorithm that averages gazepoint locations to determine the fixation location.

Some systematic error can be identified and manually removed if, when studying fixations superimposed on the visual stimuli, a consistent pattern of systematic error can be identified from trial to trial. But systematic error cannot be reduced by averaging across gazepoints. Even after calibration, eye trackers often maintain a systematic error such that fixations are recorded a small distance from the fixation point.

To accurately record gazepoints, an eye tracker must be calibrated to accommodate the physical characteristics of the human. In the commonly-used pupil-center/corneal-reflection method, for example, the eye tracker must be calibrated to accommodate the physical characteristics of the corneal bulge and the pupil, and the relationship between these features and the foveal region on the retina. The eye tracker is calibrated for each participant in an experiment by having the participant fixate a number of preset locations on the display while the eye tracker records the pupil center and corneal reflections for each location, calibrates, and then interpolates and records gazepoint locations during data collection.

Problem #2: When eye tracker accuracy deteriorates below a threshold, recalibration is necessary.

As with most instrumentation, the initial calibration may deteriorate during an experiment, at which point recalibration is necessary. Some eye tracking studies discuss this problem. One study points out that “eye tracking data is never perfect. The system may lose track of the pupil or the corneal reflection, or the observation may be simply incorrect (e.g., beyond the screen limits even though the subject is clearly looking at the screen).” (Aaltonen, Hyrskykari & Riih , 1998, p.135). These researchers addressed the problem in part by removing problematic trials from the analysis, but there are problems with this approach. Excluding more than a few trials may upset the balanced design of an experiment and limit the conclusions that can be drawn from the data.

Many studies (such as Findlay, Brogan & Wenban-Smith, 1993; Smith, Ho, Ark & Zhai, 2000) do not discuss what sort of error was observed at any point in the experiment or how accuracy was maintained or verified, making it difficult for the reader to know if these details were even considered in the experimental procedure and data analysis. Since eye tracking data contains error, this may be an important omission that limits the usefulness of the data.

A number of experimenters have addressed the calibration deterioration problem by recalibrating the eye tracker during the course of an experiment. One study addressed the problem in part by always recalibrating the eye tracker at a particular midpoint in the experiment (Aaltonen et al., 1998, p.135). Another study reports that “the subjects’ eye position was calibrated at the beginning of the session and whenever necessary thereafter.” (Pollatsek, Rayner & Henderson, 1990, p.202) In another, “an experimenter monitored each experimental trial and recalibrated the eye tracker if there appeared to be a sizable disparity between reasonable expectations about where users would look (in particular, users needed to look at the target on each trial) and the position reported by the eye tracker.” (Byrne, Anderson, Douglass & Matessa, 1999, p.404) These last two studies address the problem, but rely on the subjective perception of the experimenter to decide when recalibration is necessary, and do not specify the specific circumstances that would lead to a recalibration.

An occasional experiment (we found two, both conducted by the same primary investigator: Abrams & Jonides, 1988; Abrams, Meyer & Kornblum, 1989) is designed so that the eye tracker calibration is objectively verified before every trial, and the actual test criteria is reported in the experimental writeup. In each experiment, the participant is instructed to fixate a set of crosshairs between each trial, and the output of the eye tracker is automatically verified to be within range of this location on the display. “The calibration was performed at the beginning of each session and verified before each trial. If fixation deviated by more than 1° from the expected value, then the calibration procedure was automatically invoked.” (Abrams et al., 1989, p.536) The crosshairs served as an *explicit* required fixation location--the participant was specifically instructed to fixate the location.

An objective verification step and criteria contributes to the accuracy and usefulness of the data recorded, the confidence in the conclusions drawn from the data, and the reproducibility of the experiment. Eye movement data is inherently noisy, and eye tracker calibration deteriorates over the course of an experiment. The goal is to keep the measuring apparatus as consistent and reliable as possible. Developing and reporting procedures for addressing these problems will advance the development of behavioral research instrumentation and methodologies. To be as accurate and objective as possible, the recalibration procedure should be handled entirely by the experimental software, automatically verifying calibration between each trial, and automatically triggering a recalibration when necessary.

Many eye-tracking experiments entail tasks much more complex than looking at a set of crosshairs and then making a single saccade to a target (the task in Abrams & Jonides, 1988; Abrams et al., 1989). Eye trackers are becoming more widely used to monitor and evaluate eye movement patterns in realistic task settings, such as searching through a number of web sites looking for information (Card et al., 2001; Pirolli, Card & Van Der Wege, 2001). Asking a participant to stop periodically and fixate crosshairs to verify the accuracy of the eye tracker will sometimes disrupt the flow of the task execution and possibly the very behavior that is the focus of the inquiry.

Problem #3: Eye tracking is a software-intensive endeavor.

All told, there is a great deal of computer software involved in the full life cycle of an eye tracking experiment. Experimental software presents visual stimuli to the participant and records manual responses. Image-processing software (integrated with the eye tracking device) converts video images of the eye into gaze points on the computer screen. Fixation-detection algorithms convert gaze points into fixations (such as discussed in Salvucci & Goldberg, 2000). Software visualization tools superimpose fixations on the stimuli for analysis (such as Lankford, 2000; Salvucci, 2000). Protocol analysis tools facilitate abstract analyses of eye movement strategies (such as Card et al., 2001; Salvucci & Anderson, 2001). Whereas the image-processing software used to convert video images to gaze points is an integral component of the eye tracking machinery and isolated from the experimenter, every other piece of software is often developed or at least modified by the experimenter. For eye tracking to serve a wide range of human factors and experimental studies, more guidance is needed in the software development associated with eye tracking studies and data analysis.

THE SOLUTION

Implicit required fixation locations (RFLs) can be used to monitor eye tracker calibration in real time, and invoke recalibration when necessary. Task requirements, when combined with human perceptual and motor characteristics, often constrain where fixations will occur. In many human-device task executions, there are a number of locations that the participant must fixate in order to accomplish a task. These locations can be used, in conjunction with real-time fixation location data, to verify that the calibration of the eye tracker is still accurate.

Many task situations are designed such that the participant must fixate particular screen objects in order to accomplish a task. Some critical task information, such as a high-confidence reading of the identification code associated with an aircraft displayed on a radar screen, will only be available if the user fixates the code. Some task actions, such as clicking on screen objects, can only be accomplished if the user looks at the target.

It might be argued that because people can shift their “covert visual attention” without moving their eyes (Posner, 1980), little can be said about where the eyes will fixate. We disagree with such a position. Though covert visual attention generally shifts to the destination of a saccade before the saccade is executed, there is little or no evidence that, when people are allowed to move their eyes freely, covert visual attention moves independently of the eyes (Findlay & Gilchrist, 1998).

The target of a mouse movement, if small enough, would make a very good RFL. Eye-hand coordination research has shown that the eyes are used in the closed-loop “current control” (first discussed by Woodworth, 1899) final phase of a rapid aimed movement. Keeping the eyes on the target facilitates getting the hand or the manipulated display item (i.e. the cursor) to the target as quickly and accurately as possible. As the cursor approaches the target, the remaining distance to

the target is monitored and corrective adjustments are made until the target is reached (see Rosenbaum, 1991; Wickens, 1984). Eye movement studies confirm that, even when the hands start moving before the eyes, the eyes arrive at the target before the hand and stay on the target for the completion of the aimed movement (Abrams, Meyer & Kornblum, 1990; Helsen, Elliott, Starkes & Ricker, 1998; Smith et al., 2000). Incorporating mouse movements into an experimental task on a computer provides an opportunity for RFLs. It also contributes to the ecological validity of the task because, in real-world computer usage, a visual search is often followed by a point and click.

This article discusses a visual search study that utilizes an implicit RFL to verify eye tracker calibration between each trial. The RFL is *implicit* in that the participant is not specifically instructed to fixate the location, but must do so nonetheless to complete the next trial. Eye movement data from the experiment demonstrates that the implicit RFL procedure correctly invokes the recalibration procedure when calibration has deteriorated. The data also reveal an unexpected result--*when RFLs can be identified across the full range of a visual display, they can be used to identify and remove systematic error in the eye movement data*. This research provides guidance for developing and modifying software programs to use implicit RFLs to automatically verify eye tracker accuracy and invoke recalibration procedures when necessary, and to clean up the systematic error in eye tracking data during post hoc analysis

THE EXPERIMENTAL DESIGN WITH THE IMPLICIT RFL

We conducted a visual search experiment in which we used a single implicit RFL to validate calibration with every trial. The experiment is a study on visual search and mouse pointing in two-dimensional visual hierarchies, and replicates the experiment reported by Hornof (2001) with two exceptions: (1) the experimental procedure was modified to accommodate eye tracking, and (2) a different computer monitor was used (the pixels per horizontal or vertical cm increased from 33 to 39).

In the experiment, participants were precued with a word (or pseudo-word) and then asked to find the word in a layout such as that shown in Figure 2. Each trial proceeded as follows: (1) The computer screen is blank except for the precue. (2) The participant moves the cursor to the precue and clicks (presses and releases) the mouse button. (3) The precue disappears and the screen layout appears. (4) The participant finds the target item. (5) The participant moves the cursor to and clicks on the target item. (6) The software provides immediate and brief feedback on the performance for that trial. (7) The precue for the next trial appears.

The experiment was designed to separate visual search time from target selection time by imposing a *point-completion deadline* (Hornof, 2001). Once the participant started moving the mouse from the precue position, he or she had a limited amount of time to click on the target before the trial was interrupted. Participants were financially motivated to find the target as quickly as possible, to not move the mouse until they found the target, and once they found the target to click on the target as quickly as possible. Participants received a base payment of \$10 plus bonuses as described by Hornof (2001). Participants were allowed to spend as much time as they wanted studying the precue before clicking on it to start the next search-and-click phase. Figure 3 shows the time course of a trial, including the eye and mouse-cursor movements motivated by the experimental design.

Sixteen participants completed the experiment. An additional eight participants were screened out for one of the following reasons: calibration difficulties ($n=4$); failed stereopsis exam ($n=2$); too many trials interrupted by the point-completion deadline ($n=2$).

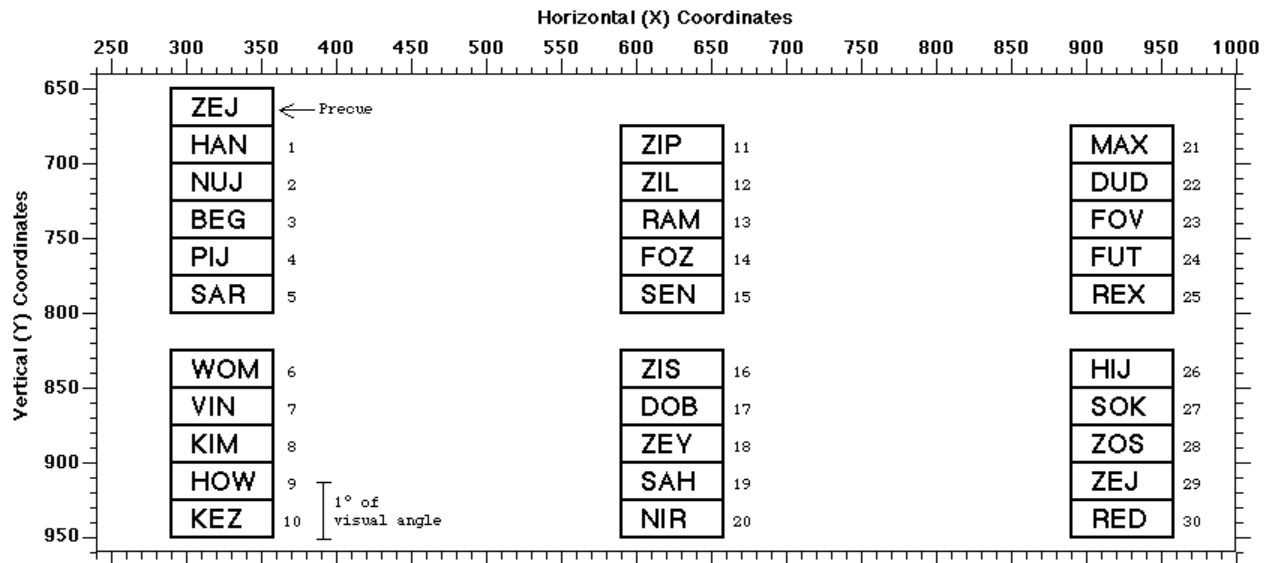


Figure 2. A screen layout with thirty items, and the screen coordinates (in pixels) where the layout would have appeared. In this trial, the target ZEJ appears in position 29. The number to the right of each screen object is the number assigned to that position. The precue is shown where it would have appeared before the layout appeared. The screen coordinate system and the smaller text did not appear on the screen during the experiment.

The experimental stimuli were displayed and mouse movements recorded using an Apple Power Mac G4 computer running Mac OS 9.2 at 733 Mhz, a ViewSonic VE170 LCD display, and an Apple Pro Mouse with an optical sensor. The experimental software was written in C++ using CodeWarrior 5. Eye movements were recorded using the LC Technologies Eyegaze System, a 60 Hz eye tracker that tracks eye movements using the pupil-center and corneal-reflection. The eye tracking system incorporates a high-speed infrared-sensitive camera fixed at the bottom center of the LCD display, an LED light source fixed in the center of the camera lens, a video frame grabber installed in a 867 Mhz Intel Pentium 3 computer running MS-Windows 2000 Professional, and various image processing and gazepoint identification software written in the C programming language. A chinrest maintained an eye-to-screen distance of 56 cm, such that 1° of visual angle subtended 38.4 pixels. The precue always appeared at the same location, level with the participant's eyes.

Each calibration and recalibration was performed using the procedure built into the Eyegaze system, as follows. Each participant sequentially fixated nine different points on the screen. The eye tracker used a least-squares regression algorithm to find an eight-coefficient calibration that optimally mapped the glint-pupil vectors to the known gazepoints. If the root mean squared error exceeded roughly 24 pixels, the system prompted for additional fixations.

The experimental design motivates some predictable oculomotor activity, which introduces a number of implicit RFLs. The first implicit RFL is the precue. It can be assumed that the participant looked at or very near the precue for each correctly-completed trial. Task analysis dictates the following activities for each trial: Look at the precue, commit it to memory, click to make the precue disappear and the layout appear, find the target, move the mouse to the target, and click on the target.

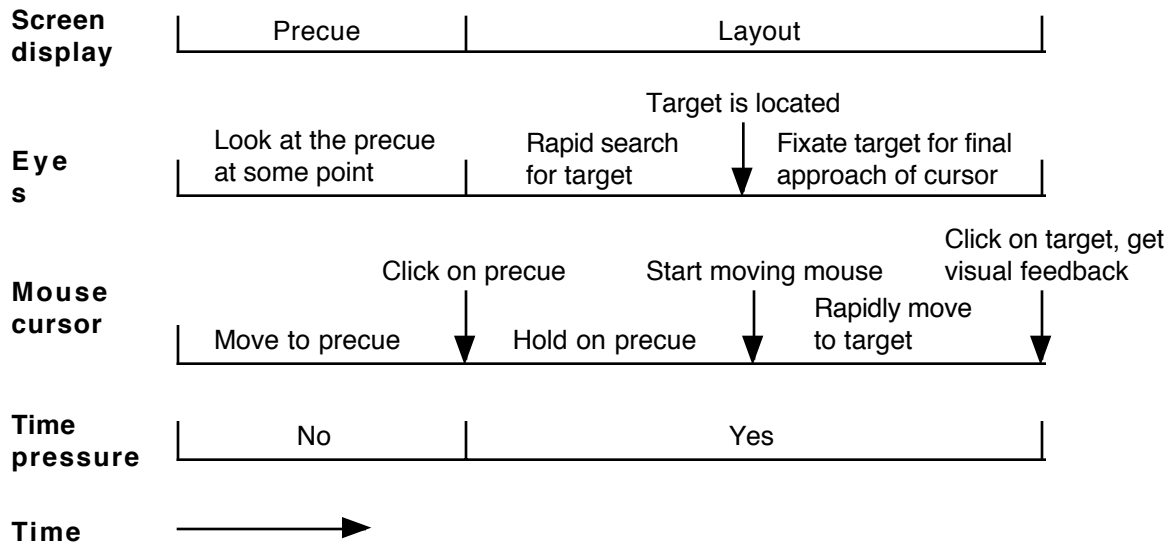


Figure 3. The time course of a single trial in the experiment, including a separate timeline for what is visible on the display, the oculomotor and manual-motor (mouse) activity motivated by the experimental design, and when the participant was pressured to perform quickly. Events are indicated with an arrow pointing to a timeline. An identical timeline for the next trial would follow immediately.

Implicit RFLs can be used to automatically invoke recalibration when necessary

The first implicit RFL--the precue--was used to verify calibration accuracy during the course of the experiment. In order to successfully complete a trial, the participant needed to look at the precue at some point while it was visible on the screen. We designed our experiment so that the eye tracker data was examined in real time during the course of the experiment. If a trial was successfully completed but not a single fixation was recorded within 2° of visual angle of the center of the precue while it was visible on the screen, then it was assumed that the participant really did look at the precue (since they knew what to look for in the layout) but that the eye tracker calibration had substantially deteriorated. At this point, immediately after the trial was completed, the experimental software automatically invoked the recalibration procedure. As well, the software marked the trial as having been conducted with a deteriorated eye tracker calibration. A new trial was generated with the same target location, and added to the pool of remaining trials.

Figure 4 shows the circular region (2° of visual angle in radius) around the precue in which a fixation had to be recorded to assume that the eye tracker was maintaining accurate calibration. We also tried setting the region to radii of 1° and 1.5° but anything less than 2° of visual angle prompted for too many recalibrations for the experiment to be completed in a reasonable amount of time. We were initially concerned about setting the threshold so high. It suggested that we would consistently find up to 2° of error in all of our eye tracking data. The error-correction technique discussed later in this article, however, explains why we had to set the threshold so high. It was not that participants were being recorded at random locations within a circular region that was 2° in radius. Rather, each participant was actually being recorded within a smaller region, and the center of the smaller region shifted within the 2° radius circle for each participant.

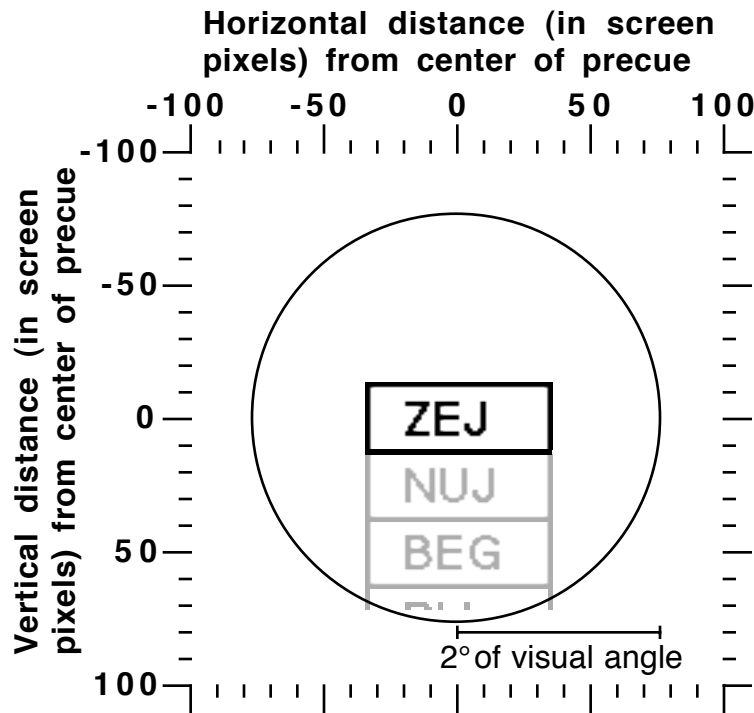


Figure 4. The precue was used as an implicit required fixation location (RFL). Though the participant was not specifically instructed to fixate it, the participant needed to view it to successfully complete a trial. The experimental software was written so that if no fixation was recorded within 2° of visual angle (77 pixels) of the center of the precue while it was visible and yet the trial was correctly executed, then it was assumed that the eye tracker calibration had deteriorated, and a recalibration was automatically invoked at the end of the trial.

The experimental software running on the Macintosh received gazepoints from the Eyegaze system in real time via an Ethernet connection. While the precue was showing, the experimental software used the gazepoints to determine fixations in real time. A dispersion-based algorithm was used to determine the center of fixations (see Salvucci & Goldberg, 2000), with a minimum fixation duration of 100 ms (6 samples from the Eyegaze system) and a deviation threshold of 0.5° radius. To accommodate noise, a gazepoint that exceeded the deviation threshold of an ongoing fixation was included in the fixation if the following gazepoint did not exceed the deviation threshold of the ongoing fixation. In addition, lost gazepoint samples (e.g. due to a blink) were assumed to be part of an ongoing fixation if fewer than 6 samples (100 ms) were lost and the gazepoint following the lost samples was within the deviation threshold of the ongoing fixation. Numerous additional technical details are available in Halverson and Hornof (2001). The experimental software and the eye tracking data reported here are downloadable at <http://www.cs.uoregon.edu/~hornof>.

The second implicit RFL is the target, which appeared in any of the thirty layout positions. It can be assumed, given our experimental design, that participants were fixating the target when they clicked on it to end the trial. Once the trial began, participants were losing money up until the moment that they clicked on the target. Once they started moving the mouse from the precue to the target, any delay in clicking on the target could result in an interruption error due to the point-completion deadline. Missing the target also resulted in an error, and either error resulted in additional financial loss. As would be necessary to perform the task quickly and accurately,

participants were motivated to get their highest resolution vision--the center of the fixation--on the target as quickly as possible, and to keep it there as they guided the cursor to the target and verified its position as they clicked. They stopped losing money as soon as they clicked on the target, so there was no motivation to look away before clicking.

Each target item can thus be used as an RFL. The target location can be compared to the fixation location recorded at the moment the mouse is clicked on the target. The distance between these two points can be used to examine how the accuracy of the eye tracker changes over time, and to evaluate the effectiveness of using the precue RFL to automatically invoke recalibration. As will be seen, it turns out that these distances can also be used in the data analysis phase to identify and compensate for systematic error in the eye tracking device.

RESULTS

The eye tracking data from the sixteen participants were analyzed to determine the effectiveness of using the precue as an implicit RFL to automatically invoke the recalibration procedure. Some confirmation could be gained by examining the fixations that were recorded during the precue phase of the experiment. Precue fixations from trials that triggered the recalibration procedure were compared to trials immediately following recalibration. Figure 5 shows two such trials. It appears as if the procedure correctly identified when recalibration had deteriorated below a threshold, and that recalibration led to a more accurate recording of fixations.

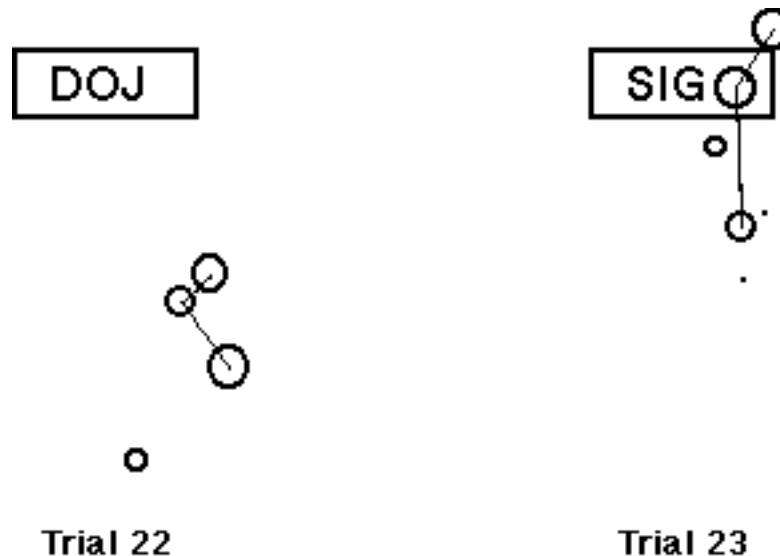


Figure 5. Fixations recorded during the precue phase for Trials 22 and 23 of Participant 3. The diameter of each circle represents the duration of each fixation, and the lines represent the order of fixations. The circles not connected by lines represent gazepoints that were not tracked, such as during a blink. Since no fixations were recorded within 2° of the precue during the precue phase of Trial 22, a recalibration was automatically invoked after the trial. Fixations appear to be recorded more accurately during the precue phase in Trial 23.

Perhaps the best evidence that the accuracy of the calibration deteriorated over time, and that the procedure correctly identified when the calibration accuracy dropped below a threshold, can be seen by examining the distance between the center of the target item and the *final fixation location*--the fixation location recorded by the eye tracker at the moment that the mouse is clicked on the correct target. As discussed earlier, the target location in this experiment can be used as an implicit RFL at this moment. Figure 6 shows the *absolute deviation*--the distance between the final fixation location and the center of the target item--for all correctly-executed trials by all sixteen participants. From the 5006 final fixations recorded, 13 outliers were removed (from this and subsequent analyses). The rest are plotted. Figure 6 also shows the best-fitting linear regression for all absolute deviations recorded between recalibrations, or the beginning or end of the session. As can be seen in the regressions, the absolute deviation tends to increase over time, and tends to be reduced after each recalibration.

The visual search experiment discussed above has two RFLs: the precue and the target. The first of the two was used to invoke an automated recalibration. This left the second RFL to be used for evaluating the effectiveness and usefulness of the automatic invocation technique based on the first.

When to interrupt for a recalibration

Implicit RFLs can be used to objectively decide when to automatically invoke the recalibration procedure. However, the exact time at which a recalibration is initiated may negatively impact an experiment. If the experiment software needs to be interrupted for a recalibration of the eye-tracker, this should be done at a point in the experiment at which the participant would not be poised to execute a time-pressured or memory-intensive response. Otherwise, the software could disrupt any motor program preparation or memorization that has been done thus far, influence the participant's decision to do such preparation in the future, and thus interfere with the very human performance that the experiment was designed to capture in the first place. Instead, the software should wait for an appropriate time in the experiment when the participant is not poised to execute a critical task.

Implicit RFLs can also be used to identify and measure systematic error

When studying the verification results, starting with those shown in Figure 6, it became evident that besides being useful for automatically invoking recalibration when necessary, implicit RFLs can also be used in the data analysis phase to measure and reduce the systematic error associated with each participant.

The absolute deviations shown in Figure 6, though confirming the procedure, suggest that there is a lot of error in the eye tracking data. The *average absolute deviation*--the average distance between the center of the target and the final fixation location--across all trials is 41 pixels, a little over 1° of visual angle. If participants were truly fixating the target when they clicked on it, which we believe to be the case, such error suggests that all of the fixations recorded during the experiment were recorded an average of 1° of visual angle away from where the participant was truly fixating. This seems slightly high, and would seem to detract from the usefulness of the data for our primary research question--figuring out how people searched the layout. Decomposing and studying the error, however, reveals that much of the error is systematic, and can be identified and removed.

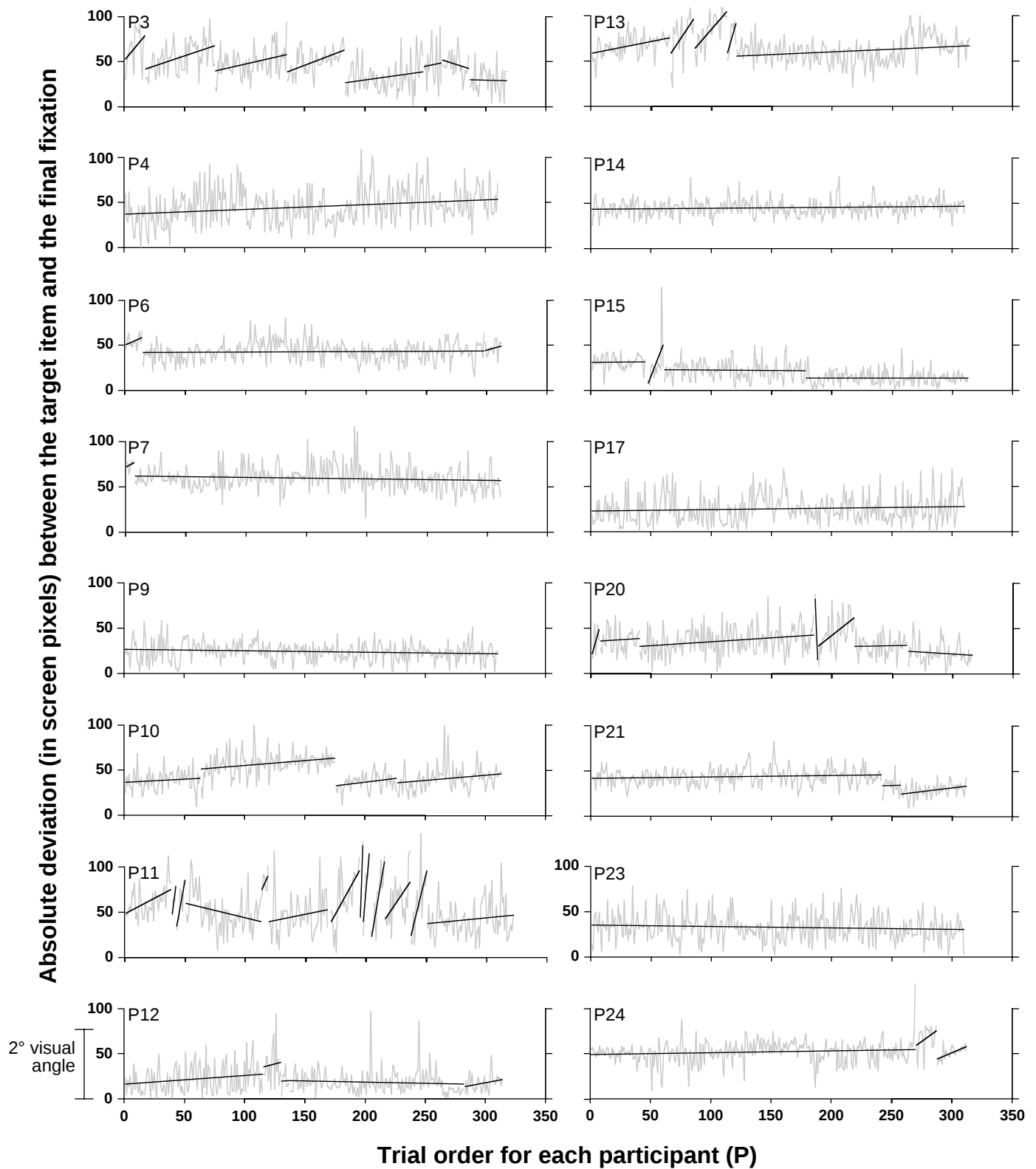


Figure 6. The absolute deviation (in screen pixels) between the center of the target object (a required fixation location--RFL) and the final fixation location (at the moment the mouse button is clicked) for every correctly-completed trial, as a function of the order of all trials. Each participant (P) is shown in a separate graph. The best-fitting linear regression is plotted for each subset of trials that occurred between automatic recalibrations (or the start or end of a session). The positive slope in most regressions suggests that the average absolute deviation tended to increase over time, and that the eye tracker calibration was gradually deteriorating for all locations, not just the precue (the RFL used to trigger the recalibration). The data suggest that comparing the fixation positions with the precue location accurately captured the calibration falling below a threshold of accuracy, and that the automatic invocation technique contributed to an overall increase in calibration accuracy across all fixations.

Figure 7 shows box plots of the absolute deviation for all participants and all trials, first aggregated across all trials, and then split out by participant. As can be seen in Figure 7, examining the absolute deviation across all participants together (the leftmost box plot) reveals that the absolute deviation is somewhat evenly distributed around 40 pixels, and that most are between 15 and 65 pixels. This rather wide distribution can be explained in part by systematic variation in the mean and the distribution of the absolute deviation from participant to participant. Participant P7, for example, is consistently recorded 45 to 80 pixels from the target while Participant P9 is consistently recorded 10 to 40 pixels from the target. Further scrutiny reveals further systematic trends.

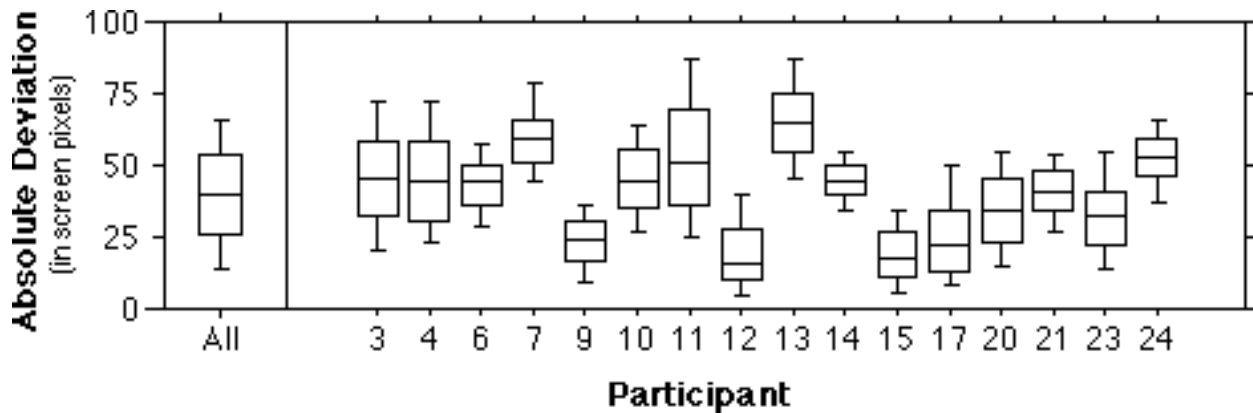


Figure 7. The absolute deviation of all final fixations for all participants (leftmost box plot) and for each participant individually. Each box plot shows the 10th, 25th, 50th, 75th, and 90th percentile. As can be seen, the median and the distribution varies systematically from participant to participant.

We have thus far examined the absolute deviation, but not the direction of the deviation. Participant P7, for example, is typically recorded an average of 60 pixels from the center of the target. But in what direction? Is he typically recorded to one side of the target, or is he 60 pixels off on all sides? To answer this question, we will next examine the *final fixation deviation*--the distance and direction from the center of the target item to the final fixation. For this analysis, we convert the distance and direction into the corresponding signed horizontal and vertical components of the deviation. A final fixation with a horizontal deviation of -50 pixels and a vertical deviation of -25 pixels, for example, would be recorded 50 pixels to the left and 25 pixels above the center of the target.

Figure 8 shows box plots of the horizontal and vertical components of the final fixation deviation, first aggregated across all trials, and then split out by participant. When the final fixations for all participants are combined (the leftmost box plot), the horizontal and vertical deviation tends to be close to zero. The average horizontal deviation is -10 pixels, and the average vertical deviation is -8 pixels, just near the top center of the target word. The large distribution in the aggregate horizontal and vertical deviations can be explained in part by individual differences among participants. As can be seen, each participant's final fixation tends to be consistently in one direction or another. All of the horizontal deviation box plots that fall below the zero grid line, for example, were consistently recorded to the left of the target center.

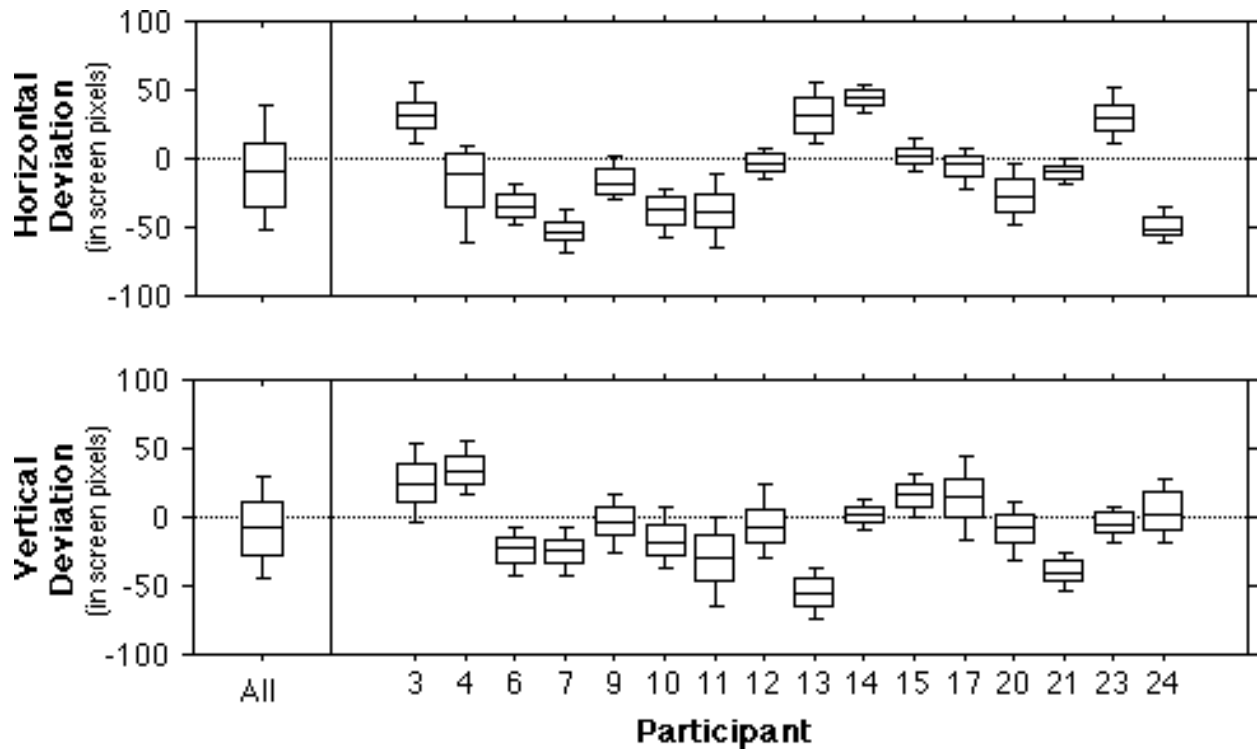


Figure 8. The horizontal (top graph) and vertical (bottom graph) components of the final fixation deviation for all participants (left frames) and for each participant individually (right frames).

Table 1 shows the means and standard deviations corresponding to the box plots in Figures 7 and 8. The mean horizontal and vertical deviations shown for each participant could be used to transform that participant's data. The standard deviations indicate the error that would not be removed by the transformation

At this point, we have identified some systematic error in the eye tracking data that can be removed by transforming the data based on each participant's average horizontal and vertical deviation. We next examine how the horizontal and vertical deviation for each participant systematically varies from position to position on the screen.

Figure 9 shows the horizontal and vertical component of the final fixation deviation as a function of the target position for two participants. An average of 10 final fixations were recorded at each target position. Each participant exhibits a rather consistent pattern of horizontal and vertical deviations across various regions of the screen and across all trials, which we refer to as the participant's *error signature*. Figure 10 shows perhaps the clearest and most useful view of the error signature--vector plots that show the mean distance and direction of the final fixation deviation for every location in the layout.

Table 1. The average final fixation deviations (with standard deviations in parentheses) for all final fixations for all participants.

Participant	Absolute Deviation	Horizontal Deviation	Vertical Deviation
All	41 (20)	-10 (34)	-8 (29)
3	46 (19)	31 (18)	24 (25)
4	46 (19)	-18 (27)	35 (15)
6	43 (11)	-34 (12)	-24 (13)
7	60 (14)	-53 (14)	-25 (13)
9	24 (10)	-16 (13)	-4 (16)
10	46 (15)	-38 (15)	-16 (20)
11	54 (24)	-38 (22)	-31 (27)
12	20 (15)	-4 (10)	-6 (22)
13	66 (16)	33 (18)	-55 (15)
14	45 (9)	44 (9)	2 (9)
15	20 (12)	3 (10)	15 (15)
17	25 (16)	-6 (12)	14 (22)
20	34 (16)	-27 (18)	-9 (19)
21	41 (11)	-10 (7)	-39 (11)
23	33 (15)	30 (16)	-5 (12)
24	53 (12)	-49 (11)	3 (20)

Note. All measurements are in screen pixels.

As can be seen in Figures 9 and 10, the final fixation deviations are sometimes consistent across the entire screen, but sometimes gradually change from one region to another. Participant P14's final fixations, for example, were consistently recorded about fifty pixels to the right of the target, but sometimes a little above the target center (at the top of each column in the layout) and sometimes a little below the target center (at the bottom of each column). Participant P4's error signature is even more interesting. The first column is recorded with an average horizontal offset of +5, the second column with an average horizontal offset of -31, and the third column with an average horizontal offset of -56. The vertical offset is smaller near the top and bottom of each column, and gradually increases for targets closer to the vertical center of each column.

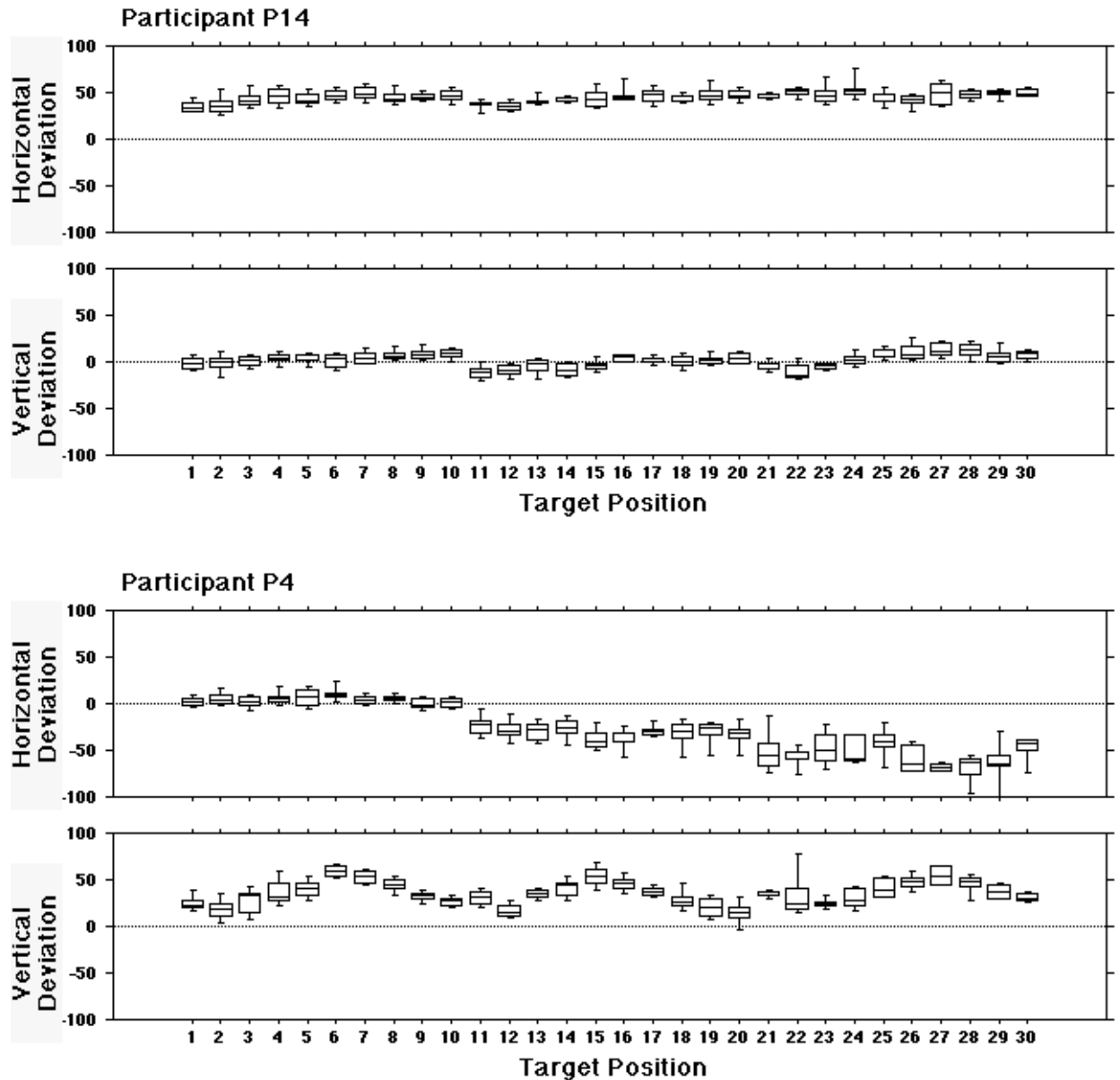
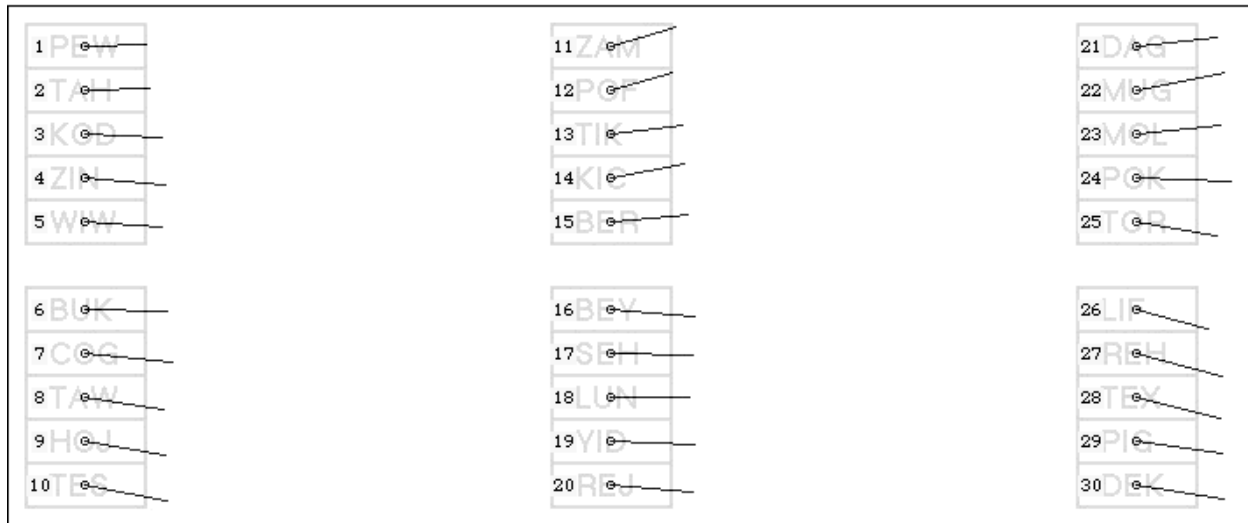


Figure 9. The error signatures of Participants P14 and P4, shown as the horizontal and vertical deviation (in screen pixels) between the final fixation locations and target centers.

We have now isolated much of the systematic error in the data, and it can be removed by transforming the data based on these error signatures. The remaining unexplained error can be measured by averaging the standard deviations of the sixty position-by-position horizontal and vertical deviations for each participant. This roughly corresponds to finding the average height of all sixty box plots shown in Figure 9 for each participant. We call this measure the *average standard deviation of the error signature*. Table 2 shows the average standard deviation of the error signature for each participant, ordered from least to greatest. Smaller measures indicate that more systematic error was successfully identified.

Participant P14



Participant P4

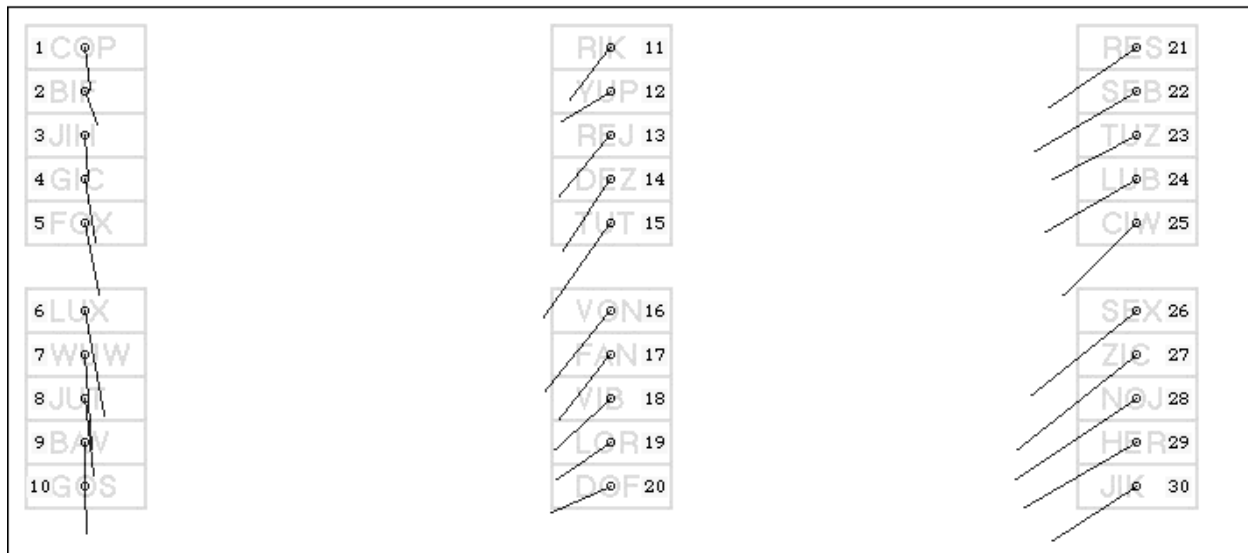


Figure 10. The error signatures of Participants P14 and P4 shown as vector plots superimposed on the thirty target positions, drawn to scale. Each vector is drawn from the center of the screen item to the average location recorded by the eye tracker when the participant clicked on a target at that location. The overall pattern of systematic error varies across all participants, but is consistent within each region for each participant. The horizontal error for P4 changes from column to column, for example, but is rather consistent within each column. These vectors were used to remove or “undo” the error in the fixation data recorded for each participant.

The average standard deviation of the error signatures shown in Table 2 fall within a relatively small range for most participants, suggesting that the error-removal technique worked well for most participants. A few participants (P3, P11, and P20) are substantially higher than the others, but note that these are the same participants that had trouble getting calibrated and tracked by the eye tracker, as can be seen by the many recalibrations for these participants shown in Figure 6. Clearly, more research is needed to further identify and remove all systematic error for all participants, but the technique described here allows us to not only remove much error for most participants, but to measure the success of the technique itself on a participant-by-participant basis, to further identify which participant's eye tracking data is likely to be more accurate and useful for general research purposes.

Table 2. Average standard deviation of the error signature for each participant, sorted from least to greatest.

Participant	Average Standard Deviation of the Error Signature
14	7.0
21	7.6
15	9.9
4	10.4
6	10.5
17	10.5
9	10.9
24	11.2
7	11.3
12	11.9
23	11.9
10	15.7
13	15.8
20	17.8
3	18.0
11	21.2
Overall Average	12.6

Studying the RFL deviations in greater and greater detail has allowed us to reduce the error in the data. An interesting question is whether we could take the analysis further, and determine how a participant's error signature changes over time--from calibration to calibration or even from trial to trial. Perhaps such an analysis would allow us to identify a different error signature each time that the eye tracker was recalibrated for Participants P3, P11, and P20. Unfortunately, it is difficult to answer these questions because, due to the variability in the eye movements and the eye tracker, a large number of trials are needed to produce smooth and useful error signatures such as those shown in Figure 10.

The systematic error observed here is in the range typically seen with our brand and model of eye

tracker. If we assume that the final location fixated by each participant was 8 pixels up and 10 pixels to the left of the center of the target item, then the average absolute deviation would be reduced from 41 pixels to 28 pixels, or 0.73° of visual angle. This is in range of the performance specifications for the Eyegaze system, which state that the system will exhibit a “typical average bias error” of 0.45° and a “maximum average bias error” of 0.70° . As stated, these “bias errors result from inaccuracies in the measurement of head range, asymmetries of the pupil opening about the eye’s optic axis, and astigmatism. They are constant from frame to frame [gaze point to gaze point] and cannot be reduced by averaging or smoothing.” (L. C. Technologies, 2001, p.8) Though they cannot be reduced by averaging or smoothing, it appears as if experimental design and data analysis such as that used here can detect and measure these constant bias errors, at which point they can be reduced in the data analysis phase. It appears as if the constant error is relatively consistent within any region of the screen for each participant, but varies across regions and participants.

Implicit RFLs can be used to remove systematic error

The error signatures reveal a pattern that can be useful in identifying and removing systematic error. Once they have been measured, the systematic error can be applied to transform the data to remove the systematic error in eye fixations that cannot be associated with an RFL, such as the fixations that occur during the visual search phase of an experiment. Figure 11 shows the fixations recorded during a trial completed by Participant P4 for a six-group layout with group labels. The figure shows (in light gray) the fixations reported by the eye tracker and (in black) the fixations after they were transformed to compensate for P4’s error signature. Each recorded fixation was transformed by cutting a horizontal and vertical line through the fixation, finding the nearest error signature vector ending in each of the four quadrants around the fixation, and adjusting the position of the fixation based on an average of the four vectors. The average was weighted based on the inverse of the distance between each vector and the fixation. As can be seen in Figure 11, the error-corrected fixations land on the screen objects likely fixated by the participant. The participant fixated a number of group labels, landed on *X4X*, searched within the target group, and ended on the target *RUB*.

The error between the center of the target item (an implicit RFL) and the recorded final fixation location reveals a unique and consistent pattern of error for each participant. When RFLs can be identified—that is, when it is very likely that a fixation occurred at a particular location at a particular time—the disparity between the observed fixation location and the RFL can be recorded. If these disparities are collected across the entire scene for each participant, a unique error signature can be identified for each participant. The error signature can be used to remove error in fixations throughout the entire visual scene, and not just when the fixation was on an RFL.

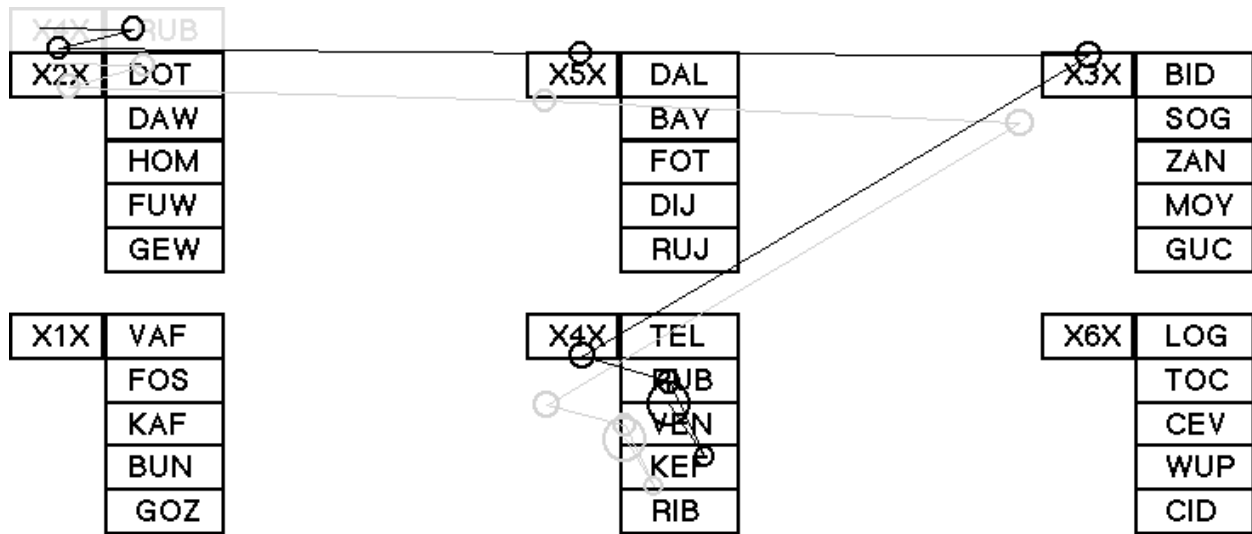


Figure 11. The fixation data recorded by the eye tracker (light gray circles) and after the error signature transformation is applied (black circles). Participant P4 searched a six-group layout with group labels. The diameter of each circle indicates the duration of each fixation. The precue X4X RUB is shown in the top left, where it would have disappeared when the layout appeared. The participant first searched the group labels until she found the group X4X, and then confined her search to that group. The final fixation is on the target RUB. It appears as if the RFL-based error correction successfully removes much of the systematic error in the data, and puts the recorded fixations closer to the actual fixations made by P4. Note how the error correction varies from region to region and corresponds to Participant P4's error signature in Figure 10.

Task analysis must be used to identify RFLs. It cannot always be assumed, for example, that a participant will be looking where he or she is clicking. The task parameters of the experiment described here specifically motivated participants to look at the target when they clicked on it. Given other task parameters, participants might sometimes position the cursor on a target, move the eyes away from the cursor, and then click. As well, if detailed task-critical information will be visible during a period of time, the experimenter cannot necessarily predict exactly when the information will be fixated, but only that a fixation will fall near that location.

It is possible that error measurement conducted with RFLs as discussed here could be integrated with the automated eye-movement protocol analysis techniques developed by Salvucci and Anderson (2001), to further improve the overall usefulness of one or both techniques. In the protocol analysis, hidden Markov models (HMMs) are used to classify patterns of fixations into search strategies. The straightforward integration of the two techniques would be to first remove the error by compensating for measured RFL deviations, and to then apply the protocol analysis. But it is possible that the two methodologies could be integrated. For example, in the HMMs, there are a number of low-probability states labeled "any" in which the eye could be looking anywhere or in which there was too much error to identify the fixation. A small percentage of fixations may be attributed to these states and the overall pattern of eye movements still classified as belonging to a particular strategy, provided that enough other fixations can be attributed to states that correspond to specific locations. Perhaps the RFL deviations could identify areas of the display in which the variance in the systematic error increases, and the probabilities for the "any" states that are yoked to expected fixations in those regions of the display could be increased slightly.

CONCLUSION

In an eye tracking experiment, one computer system or subsystem is typically used to present the stimuli to the participant, and another to collect the eye movement data, with little interaction between the two during the course of the data collection. This paper demonstrates some of the benefits of having the two systems--the eye tracking system and the stimulus-presenting system--interact during the course of an eye tracking session. In this paper, the interaction leads to the recording of more accurate data during the experiment, and more accurate data analysis after the experiment. The general pattern of having the two systems interact could be repeated for many other uses, such as changing the experimental design mid-session based on eye movement patterns observed earlier in the session, or for multi-modal interfaces that respond to eye movement patterns in real time, such as to alter the reliance on the eye movement data in real time.

An experimental design will motivate and perhaps even dictate patterns of eye movements. A careful consideration of the interactive system design and human performance characteristics can be used in tandem with the eye tracking data--both in real time and in the *post hoc* data analysis--to use an eye tracker for varied and innovative applications, for both experimental purposes and real-time applications. The two systems--the stimulus-presentation system and the eye-tracking system--can communicate with each other in real time to facilitate a rich set of experimental designs and applications, and to record more accurate eye tracking data, and to clean up systematic error in eye tracking data by using required fixation locations.

ACKNOWLEDGMENTS

The authors would like to thank Ron Chong, Dixon Cleveland, Maureen Ford, Wayne Gray, Luke Hornof, Art Joyce, Pete Norloff, Dario Salvucci, Mike Schoelles, Jeffrey Shell, Lawrence Toto, Jonathan Vaughan, and one anonymous reviewer for discussions and feedback on this research. This work was supported by the Office of Naval Research through Grant N00014-01-10548 to the University of Oregon, Anthony J. Hornof, principal investigator.

REFERENCES

- Aaltonen, A., Hyrskykari, A., & Räihä, K.-J. (1998). 101 spots, or how do users read menus? *Proceedings of CHI 98*, New York: ACM, 132-139.
- Abrams, R. A., & Jonides, J. (1988). Programming saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3), 428-443.
- Abrams, R. A., Meyer, D. E., & Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), 529-543.
- Abrams, R. A., Meyer, D. E., & Kornblum, S. (1990). Eye-hand coordination: Oculomotor control in rapid aimed limb movements. *Journal of Experimental Psychology: Human Perception & Performance*, 16(2), 248-267.
- Byrne, M. D., Anderson, J. R., Douglass, S., & Matessa, M. (1999). Eye tracking the visual search of click-down menus. *Proceedings of CHI 99*, New York: ACM, 402-409.
- Card, S. K., Pirolli, P., Wege, M. V. D., Morrison, J. B., Reeder, R. W., Schraedley, P. K., & Boshart, J. (2001). Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. *Proceedings of CHI 2001*, 498-505.
- Chapanis, A. (1949). Theory and methods for analyzing errors in man-machine systems. *Ann. N. Y. Acad. Sci.*, 51, Article 6.
- Findlay, J. M., Brogan, D., & Wenban-Smith, M. G. (1993). The spatial signal for saccadic eye

- movements emphasizes visual boundaries. *Perception & Psychophysics*, 53(6), 633-641.
- Findlay, J. M., & Gilchrist, I. D. (1998). Eye guidance and visual search. In G. Underwood (Ed.), *Eye Guidance in Reading and Scene Perception*. Amsterdam: Elsevier, 295-312.
- Halverson, T., & Hornof, A. J. (2001). *Issues in the design and development of experimental software for use with an eye tracking system*. Department of Computer and Information Science Technical Report, University of Oregon, downloadable at <ftp://ftp.cs.uoregon.edu/pub/hornof/EyeGazeMac.pdf>.
- Helsen, W. F., Elliott, D., Starkes, J. L., & Ricker, K. L. (1998). Temporal and spatial coupling of point of gaze and hand movement in aiming. *Journal of Motor Behavior*, 30(3), 249-259.
- Hornof, A. J. (2001). Visual search and mouse pointing in labeled versus unlabeled two-dimensional visual hierarchies. *ACM Transactions on Computer-Human Interaction*, 8(3), 171-197.
- L. C. Technologies. (2001). *The Eyegaze Development System: A Tool for Eyetracking Applications*. Fairfax, VA: LC Technologies, Inc.
- Lankford, C. (2000). Gazetracker™: software designed to facilitate eye movement analysis. *Proceedings of the Eye Tracking Research and Applications Symposium*, New York: ACM, 51-55.
- Pirolli, P., Card, S. K., & Van Der Wege, M. M. (2001). Visual Information Foraging in a Focus + Context Visualization. *Proceedings of CHI 2001*, 490-497.
- Pollatsek, A., Rayner, K., & Henderson, J. M. (1990). Role of spatial location in integration of pictorial information across saccades. *Journal of Experimental Psychology: Human Perception and Performance*, 16(1), 199-210.
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 3-25.
- Rosenbaum, D. A. (1991). *Human Motor Control*. New York: Academic Press.
- Salvucci, D. D. (2000). An interactive model-based environment for eye-movement protocol analysis and visualization. *Proceedings of the Eye Tracking Research and Applications Symposium*, New York: ACM, 57-63.
- Salvucci, D. D., & Anderson, J. R. (2001). Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16, 39-86.
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocol. *Proceedings of the Eye Tracking Research and Applications Symposium*, New York: ACM Press, 71-78.
- Smith, B. A., Ho, J., Ark, W., & Zhai, S. (2000). Hand eye coordination patterns in target selection. *Proceedings of the Eye Tracking Research and Applications Symposium*, New York: ACM Press, 117-122.
- Wickens, C. D. (1984). *Engineering Psychology and Human Performance*. Columbus, Ohio: Charles E. Merrill Publishing Company.
- Woodworth, R. S. (1899). The accuracy of voluntary movement. *Psychological Review*, 3, Monograph Supplement No. 2.