How not to Use eye tracking to identify learning

Wayne J Ryan
Computer Science
Clemson University, Clemson, SC 29634
wryan@clemson.edu

ABSTRACT
It was hoped that the results of this experiment could be used to aid in the development of effective educational software. I made many experimental errors, and was therefore unable to obtain significant results. In this paper I describe my experiment and the mistakes in an effort to show what I learned. I hope that the reader of this paper will gain ability to understand and avoid the same mistakes.

Author Keywords
Eye Tracking; Learning; Education; Audio, lessons learned

INTRODUCTION
The goal of this study was to gather eye-tracking data during an educational presentation, and use the data to predict what a participant will remember.

Writing my own calibration routine seemed like a good idea. Though it was a good academic exercise I could have saved valuable time by modifying existing code. I did not verify calibration until after the tests were preformed, and therefore had to throw away much data.

Participants need the opportunity to try out the eye-tracker. Many participants need detailed instruction on where to position their heads, and practice focusing on calibration dots. I underestimated the need to "train" participants.

I used a pretest/posttest design, but I made the posttest too easy and was therefore unable to get significant results. An early pilot study may have revealed this problem before it was too late.

I encountered problems with the audio library that I chose. The problem still has not been resolved. It may be the result of a bad installation of the library files, hardware conflicts, or conflicts with the operating system. This problem was detected late because I developed the software on my laptop (it works great on my laptop). I didn't test it on the computer used in the experiment until too late.

EXPERIMENTAL DESIGN
This was a between subjects pretest-posttest design. The test condition included audio with the presentation while the control group had no audio. The pretest identified what the participant already knew (or didn't already know). The layout of the presentation was fixed, but the content was determined by the pretest.

The reasoning for altering the content was to compensate for prior knowledge. For all participants the presentation and posttest only contained questions answered incorrectly in the pretest. The posttest was intended to measure what the participant remembered from the presentation. I expected the eye-tracking data to show that participants remembered what they looked at, and didn't remember things not look at.

The audio/no audio condition was more of an afterthought and allowed me to call this an experiment rather than an exploratory study. I expected that people would look at what they heard while they heard it. The participants with no audio were predicted to look around more randomly.

I suspect that the more systematic visual pattern induced by the audio would be more conducive to learning. It has been shown that a systematic scanpath is better for visual inspection [1].

PROCEDURE
The experiment was made up of four computerized parts. They are as follows: calibration, pretest, presentation, and posttest.

Calibration lasted about 1 minute. The participants were asked to sit relatively still because the eye-tracker can't compensate for large amounts of head movement. They were instructed to sit comfortably and click the screen when ready. The computer would then display four red dots in sequence. They were asked to look at the dots.

The pretest consisted of a series of multiple choice questions. For each question a 120x120pixel image was displayed centered on the screen. Below the picture six Spanish phrases were displayed. One of the six phrases related to the image. See Figure 1. The participants were asked to click on the phrase that corresponded most closely

with the image. Additional questions were presented until nine had been guessed incorrectly.

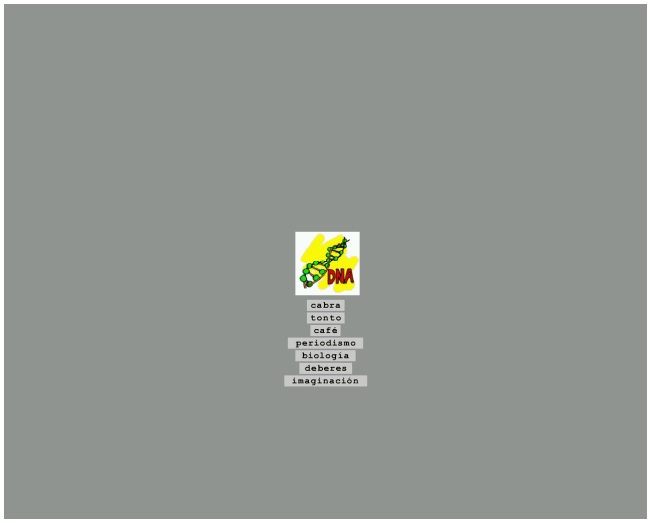


Figure 1 (example test question)

The presentation was made up of the nine word image pairs that had been guessed incorrectly. All nine pictures were displayed, randomly ordered, in a 3 by 3 grid pattern. Each picture had a corresponding phrase displayed underneath. The participants were instructed to remember which word went with each picture. They were instructed to do nothing with the keyboard or mouse during the presentation simply look, try to remember, and wait for the presentation to end. The presentation lasted about 1 minute. Eye-tracking data was collected during the presentation.

The audio presentation was identical to the no audio except that the participants heard a recorded voice say the words in order from left to right, top to bottom.

The posttest was like the pretest. It consisted of multiple choice questions. Each question included a 120 x 120 pixel image centered on the screen, and 6 phrases in Spanish below the image. There were nine questions in the posttest. The posttest questions included the same images as the presentation. Participants were asked to click on the word they remembered from the presentation.

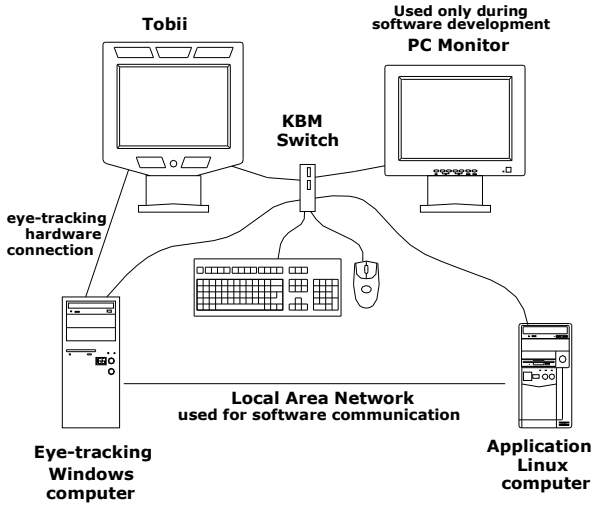
All instructions were given to the participants prior to calibration. They were told just enough so that they would know what to do. They were not told details of how the images for the presentation and posttest were determined.

EQUIPMENT

A Tobii 1750 eye-tracker was used. The Tobii is a video based, NIR corneal reflection eye-tracker. The Tobii has a 17" LCD with 1280x1040 screen resolution. Built into the monitor is a camera with optical filters, and near infra-red light-emitting diodes[2]. The eye-tracker server software uses a PC running Microsoft windows. The Windows machine is connected by LAN to a client computer running Linux RedHat. Both computers are connected to the same monitor, keyboard, and mouse. A simple KBM switch allows a user to alternate between machines. The software written for this project ran on the Linux computer and controlled the display while the Windows machine handled the eye-tracking hardware.

Screen coordinates of gaze points are calculated by the Tobii. The sample rate is approx 50 Hz. The Tobii also supplies an integer indicating validity of each gaze point. This integer is in the range from 0 to 4. If the validity is 0 then (with proper calibration) we can be very confident that the gaze point is valid. A validity of 4 would indicate that the Tobii was unable to detect the participants' eye during that sample cycle. Possible causes of invalid data could include blinks, head out of range, or rapid head movement. Validity data returned by the Tobii does not account for calibration error. A poorly calibrated machine can return inaccurate gaze-points with a validity of 0. I discarded all data with validity other than 0.

When the Tobii is properly calibrated it is accurate to within 1 degree of eye rotation.



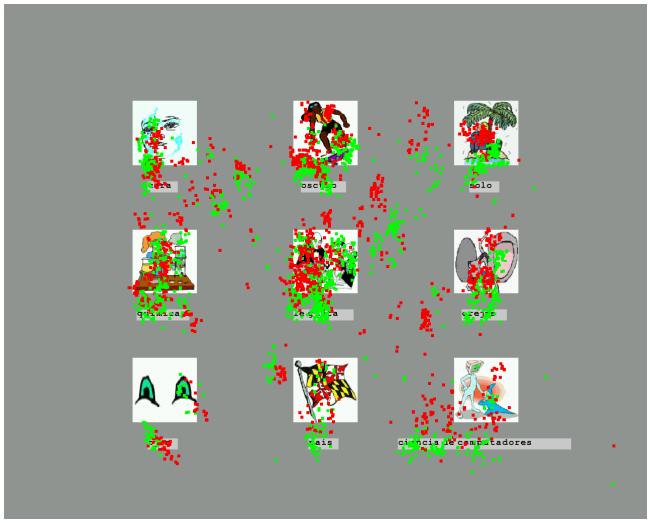


Figure 2 (good calibration)

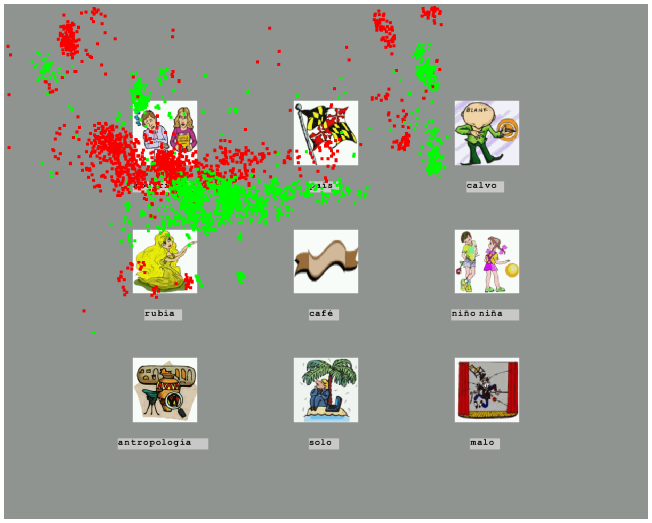


Figure 3 (bad calibration)

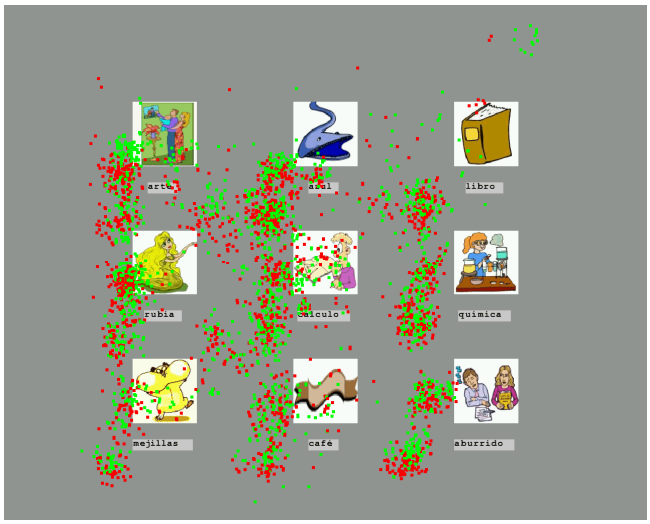


Figure 4 (before correction)

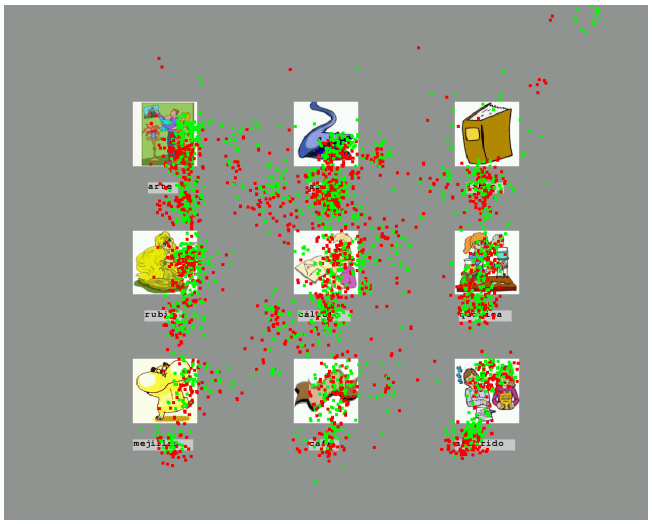


Figure 5 (after correction)

DATA ANALYSIS

The first step was to identify and isolate usable data. I wrote my program in such a way that it output a screenshot with the raw gaze-point information smattered on top. Red dots represent gaze points for the right eye, green dots for the left. By visual inspection of these screenshots I threw out 4 tests because of what I believe to be bad calibration. I threw out the first test because the computer program was not set up correctly. See Fig 2 for example of a good calibration, Fig 3 for example of a bad calibration. Not all calibration errors resulted in loss of data. If the data was clearly off by a constant factor I adjusted it by shifting all data points. The amount of shift was determined visually. See Fig 4 and 5 for an example of a test before and after correction.

After eliminating tests with calibration errors I had to distinguish fixations from saccades. A fixation is when the eye is relatively still. A saccade is when the eye rapidly moves between fixations. It is commonly accepted that people only absorb visual information during fixations[3]. It is also accepted that fixation duration is related to level of cognitive activity. For this reason saccadic data was of little interest. There are many popular methods of fixation detection each with benefits and drawbacks[4].

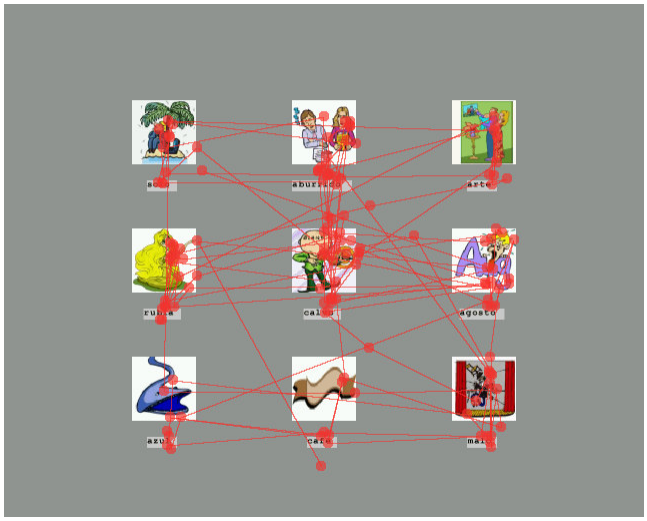
I chose velocity based fixation detection because of its ease of implementation. The theory is quite simple. The eye velocity during saccades is greater than during fixations therefore calculate the eye velocity. Determine a threshold, and label anything above the threshold a saccade, and anything below a fixation. Velocity is calculated by taking the distance between consecutive points and dividing by the time elapsed. Usually this is represented in angular velocity. It can be assumed that the eye is at a distance of 50cm from the screen. The formula for calculating angular velocity is as follows.

$$\alpha = \tan^{-1}(\sqrt{\Delta x^2 + \Delta y^2} / d)$$

Where d is the distance of the participants' eye from the computer screen.

In practice velocity based fixation detection can be quite sensitive to instrument noise [4] (random bad data points caused by machine error). To mitigate this I used a 5-tap FIR filter [5] and averaged the gaze point locations for the right and left eye. When using only point to point velocity the threshold may need to be inferred due to aspects of data collection and exploratory data analysis[4]. Normal fixation durations typically last between 100 and 500 ms [4]. I used 4 degrees/second. This gave me fixation durations in the proper range. It does however seem like a rather low threshold.

Once the points are labeled as fixation or saccade, consecutive fixation points are averaged together and fixation durations are calculated. All points labeled as saccades were removed from the data set.



I wrote a visualization program which enabled me to replay each presentation while dynamically displaying dots for each fixation point. The above screenshot shows this visualization with all data for one test. The value of this visualization is that I was not only able to see the locations of fixations but also the order in which they occurred and their duration. With more time more elaborate visualizations could be created. For example size of the fixation dot could be relative to the variance in gaze points used to calculate the location. I also would like to have adjusted the transparency of fixation dots to represent the density of gaze points. Note that density and variance are not quite the same. Two fixations can have the same variance but different density if the gaze point quantities are different.

Total dwelltime per ROI was calculated by totaling the durations of all fixations within the ROI. The each ROI was then put into one of the four categories. These dwelltimes for these four categories are compared in Chart 1.

Audio	Age	Gender	Education	Eye-tr. Exp.	Span Exp.	Multi. Ling.	No. correct on pretest	No. errors on posttest
no	26	M	S	No	None	No	7	1
no	21	F	F	No	1 yr	No	12	2
no	22	M	G	Yes	.5 yr	No	27	4
no	22	M	G	No	1.5 yr	No	10	0
no	20	F	S	No	None	No	8	1
no	22	M	G	No	1 yr	No	13	0
no	19	F	S	No	1.5 yr	No	37	0
no	37	M	G	No	1 yr	No	2	8
yes	24	M	G	No	.5 yr	No	6	0

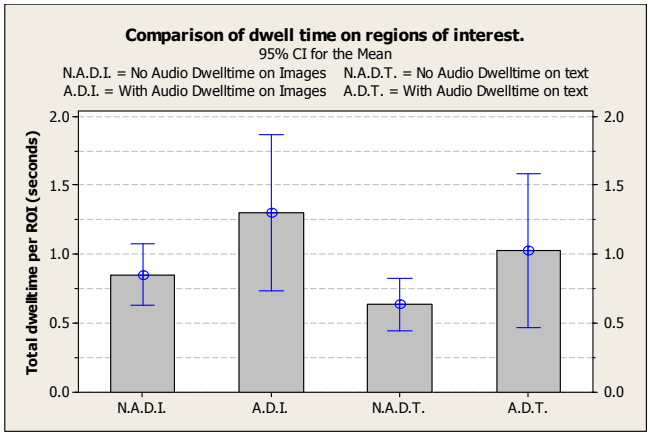


Chart 1

RESULTS

I ran 14 tests with 11 male and 3 female Clemson University students. The average age was 23. They had on average 1 year of education in Spanish. None had more than 2 years of Spanish training. None were multilingual.

I had to discard 5 tests leaving only one usable test with audio. The table above summarizes the participants' general information.

The post test was too easy. Ideally the participants will get on average 50% of the questions on the posttest incorrect. The distribution of scores was too skewed to be of much use. It is interesting to note that all five participants who heard audio made no mistakes on the posttest.

Dynamic display of fixation points leads me to suspect that audio has an impact on what people look at. Without audio people generally tend to look at the images in random order. With audio people tend to look at the image most closely related to the audio. This theory would make sense, but I simply did not have enough valid test data to support this it with any significance.

DISCUSSION

As you can see I made many mistakes.

Instruction given to the participants was not adequate or entirely consistent. One participant said "During the presentation I didn't know what to do." The participants didn't need to do anything during the presentation, just look at the screen. A pilot study could have given me a chance to understand what I needed to say, and allowed me to rehearse so that I could have been more consistent.

I had some difficulty finding participants. Most were given instruction individually. I found a group of five students all willing to participate. I tried to explain the process to them all at one time. None of them calibrated correctly. Apparently when I talked to them as a group they paid far less attention. I now realize that a small difference in the instructions given can cause a significant difference in the results.

While testing my program I created a switch that would allow dynamic display of the gaze point data. With this setting on I could watch the dots flicker to different positions on the screen as I moved my eyes. It helped me to verify that my program was activating the Tobii properly. It was easy to see the difference between a good and bad calibration. I accidentally left this switch on for the first test subject. The data collected from that participant (even though I had to throw it out) looked better than most. If I do another eye-tracking experiment I might consider letting the participants practice calibrating the machine and looking around with a similar type of display so that they can get used to the calibration process.

Calibration procedures lacked the proper verification. The "canned" software provided with the Tobii includes a feature that allows the user to look at a graphical representation of data collected during calibration. The user can then choose to continue or recalibrate. Had I incorporated a similar feature into my own calibration routine I could have avoided loss of so much data.

My software has bugs. The most glaring bug is that the audio didn't function properly. It sounded "garbled" one person described it as "being down-sampled too much". This was due to inexperience on my part. I had never used an audio library before. Whenever trying something we've never tried before we can expect to encounter some unexpected problems. The only solution I see is "test early, test often."

The Posttest was too easy. Participants got nearly all questions in the posttest correct. One way to increase difficulty would be to ask participants to recall rather than just recognize. For example they could have been asked to type the word rather than select it from a list. Recall is more difficult than recognition. I could also have measured response time (how long it takes to answer a question). It would have been interesting to compare response time to dwell time on ROIs.

ACKNOWLEDGMENTS

Andrew Duchowski, Prof. for course

Mónica Cecilia Muñoz Torres, recorded Spanish

REFERENCES

1. Sajay Sadasivan, Joel S. Greenstein, Anand K. Gramopadhye and Andrew T. Duchowski Use of Eye Movements as Feedforward Training for a Synthetic Aircraft Inspection Task: *CHI 2005*
2. User Manual : *Tobii eye-trackerClearView analysis software*

   http://www.tobii.se.
3. Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke and Geri Gay. Accurately Interpreting Clickthrough Data as Implicit Feedback. *SIGIR'05, August 15–19, 2005, Salvador, Brazil.*
4. Dario D. Salvucci, Joseph H. Goldberg. Identifying Fixations and Saccades in Eye-Tracking Protocols: *Proceedings of the Eye Tracking Research and Applications Symposium 2000.*
5. Andrew Duchowski, Eric Medlin, Nathan Cournia, Hunter Murphy, Anand Framopadhye, Santosh Nair, Jeenal Vorah, and Brian Melloy. 3D Eye Movement Analysis: *Behavoir Research Methods, Instruments, and Computers2002*