Considerations for Using Eye Trackers during Usability Studies

Anjali Phukan and Margaret Re

UMBC, Baltimore, MD, USA
{anjai2,re}@umbc.edu

Abstract. The purpose of this usability study was to see if eye trackers collect valid data, regardless of the user's method of corrected vision, eye color, or gender. The motivation to explore the idea that these human factors can distort eye trackers is based on marketing claims by several companies that say these factors should not affect results. This study found that the validity of data in usability studies that involved eye trackers in testing can produce biased results based on eyewear and eye color, and that adjustments should be made to control for these variables. The results showed no significant correlations based on gender. As a consideration into developing international signage for mass transportation systems that effectively accommodate global users, this study also explored how first language affects the way in which a user views and organizes a message and hence interprets procedural directions and related imagery. This is within the context of usability testing for a wide variety of users who may not share a first language or have the same method of vision correction.

Keywords: eye trackers, usability testing methodologies, internationalization, eye color, eyewear, gender, and language.

1 Introduction

Some eye tracking hardware manufactures make claims about the validity of their systems in terms of the demographics of the users. However, if these claims are not accurate, then researchers may design studies incorrectly, and hence report incorrect results. This initial study looks at the validity of some of these claims, and attempts to suggest usability methods to account for any unexpected variances in data that may occur as a result of these demographics.

2 Motivation

Several eye tracking hardware and software makers said any eyewear is tolerable. For instance, Tobii says "…Tobii T60 and T120 Eye Trackers track basically everyone, regardless of ethnic origin, age, glasses or contact lenses..."[1] In addition, LC Technologies, makers of Eyegaze systems says "…The system does not get confused by reflections off glasses...or by bright or dark facial features." [2]. This study is focused on seeing if eye tracking systems provide different results for users of different

N. Aykin (Ed.): Internationalization, Design, LNCS 5623, pp. 301–307, 2009.

demographics (i.e. eye color, eyewear, gender, and first language), based on these claims. These study also will see if and how the results of the different demographic groups may impact usability studies. This could have implications in many fields, from testing readability of a font [e.g. 3] to user interaction research [e.g. 4]. For more on eye tracking systems, methodologies, and other applications, see [5].

3 Study Design, Assumptions, Definitions

This study used a Tobii T120 and related software Tobii Studio to collect data on html pages with text and jpeg images. There were several Area of Interest (AOI) sections on each web page. In addition to analysis on an AOI, overall analysis on the data looked at general user trends based on various demographics. The analysis includes data gathered from the following:

- Eye movement: eye gaze validity, fixation data, pupil size
- AOI data: count, length, time to each first AOI fixation
- Other user actions: Mouse clicks, Keystrokes, URL starts and stops
- User demographics, gathered at the beginning of the study

The study collected demographic and eye gaze data in a lab setting using predefined web pages of jpeg images and 2 customized directions pages. This was done in two sets of data collection. The first set of data collected was via a pilot study of 9 graduate (Information Systems/Human Centered Computing) students and 1 teacher reviewing 8 jpeg images. The graduate students, who were MA and PhD students, with a variety of undergraduate degrees, viewed informational images specifically designed to show hierarchy of information. The study process was modified so that the survey was administered completely using paper, to remove all possible complications once the user began to use the eye tracking computer. The two studies used the same image viewing directions but the instructions were provided differently, the surveys were completed differently, and the users viewed different images.

The second set of data collected was via a second study of 25 undergraduate students viewing 4 new jpeg images. The undergraduates, senior design students, viewed several visual dictionaries that other classmates had created. The first and last pages of each study session were directions written in H1 and/or H2 on a simple hypertext markup web page. Each visual dictionary was a combination of numerals and letters of a particular typeface (e.g. Gill Sans, Akzidenz-Grotesk) in a jpeg image, where the angle, color, and size of a letter varied based on the intent of the visual dictionary's designer(s). This was a within subjects study, hence all participants viewed all visual dictionaries.

User demographics in this study group, such as gender and eyewear, are shown in Table 1. Second Study User Demographics. Users determined their own 1st alphabet, eyewear, eye color, and gender, although some users asked for advice from other participants or the researchers about the eye color. There were no noted discrepancies from what users picked as their eye color and what the observers felt was their eye color. Eyewear was based on what the user was wearing during the study, whether or not they usually use that type of eyewear, although all participants said they had normal-to-corrected vision with whatever eyewear they were using during the study. All

participants in both studies spoke English, but some users in each group used a non-Latin Alphabet as a native language. In the second study group used in this paper, there were 4 participants whose native alphabet were non-Latin, including 3 Korean and 1 Slovak student. All participants who used a Latin based native alphabet learned English as their first language, although their countries of origin varied slightly.

Table 1. Second Study User Demographics

1st Alphabet	Eyewear	Eye Color	Gender
Latin/Other	Contacts/Glasses/None	Blue/Brown/Green/Hazel	Female/Male
21/4	6/6/13	10/7/3/5	17/8

4 Findings and Conclusions

There were findings are based on the second study discussed in the previous section. The results are discussed as follows: eye color, eyewear, gender, and lastly language.

4.1 Eye Color

Larger pupils users have less valid data. Larger pupil sizes are not significantly correlated with a language or eye color, except for hazel eyes that tend to appear to have smaller pupils. In addition, glasses gather the least amount of valid data, followed by contacts, and then no corrective eyewear. Less valid data as a result of these conditions affects correlation tables, and could affect the analysis and interpretations of studies. None of these factors impacted a user's ability to navigate the tasks.

The software used to gather fixation information first verifies a piece of information via a proprietary formula that populates a validity field in a user/session table of data collected. The score ranges from 0 to 4 where 0 is valid and 4 is invalid and are significantly correlated with the pupil size of the eye being validated. The left eye pupil size was correlated to the left eye validity code a rate of .945, and the right eye's pupil size to validity code correlation was .996. When left and right eye data was combined however, the total correlation for pupil size to validity was only .88. This is because, although a piece of eye tracking data can have information on both eyes. There is a tendency for the data to be valid on only one eye at a time, where the pupil size of the left eye was significantly negatively correlated with the pupil size of the right eye. This left/right pupil size correlation and the left/right validity correlation were both valued at -.77. One thing to note is that the validity increases as the data becomes more invalid, so that the positive correlation really means that the smaller pupil tends to be more correlated with more valid data.

Hazel eyes overall did not have significantly different pupil sizes. However, in terms of left eye or right eye, they were clearly different. This perceived pupil size difference impacted the validity of the data collected in regards to left and right eye data, although overall the amount of data gathered was only slightly more than that from other colored eyes. It's interesting to note that hazel eyes are caused by a combination of a moderate amount of melanin in the iris' anterior border layer and

Rayleigh scattering (scattering of light or electromagnetic radiation by particles much smaller than the wavelength of light - can occur when light travels in transparent solids and liquids, such as gases) [6]. This could be the reason for this difference in pupil size during data collection, although there are many other possible reasons. The end result is less data collected for hazel eyes, possibly biasing the data collected, especially the fixations per minute, or even area of interest in a marketing research study, as is shown in Table 2. Average Fixations by Eye Color.

Table 2. Average Fixations by Eye Color

Eye Color	Average Fixation Duration (std)	Average Fixations per minute (std)
Brown	590.15 (132.64)	170.82 (19.87)
Blue	661.85 (84.13)	154.38 (16.48)
Green	661.90 (152.35)	147.93 (13.83)
Hazel	651.19 (101.89)	125.42 (27.69)
All Eyes	639.65 (108.29)	152.42 (24.47)

Table 3. Correlations by Eye Color shows the correlations of color to the validity code of all data gathered from an eye tracker. It is also partially correlated on the colors that had some significant effect on the validity code or average fixations per minute, where '*' indicates statistical significance at .05 and '**' indicates statistical significance at .01. The data gathered did not show an affect of eye color on average fixation duration. The results do not conclude that the eye color plays a role in average fixations or in validity of the data, but do show that there is some correlation, where darker eyes perform better than hazel eyes.

Table 3. Correlations by Eye Color

	Eye Color	Brown	Hazel
Left Eye Validity	(*)-0.449	0.187	(**)-0.548
Right Eye Validity	(**)0.537	-0.248	(**)0.644
Average Fixations per minute	(*)-0.442	(*)0.479	(**)-0.563

4.2 Eyewear

The number of fixations is correlated to a user's eyewear, where glasses include both reading glasses and every day glasses. There were no users of bi-focal glasses in this study. In addition, eyewear in general loosely correlated with 7 of 9 AOIs on the directions page of the study task, where users not wearing contacts or glasses had the higher fixation counts, fixations lengths. On the image pages, users with corrected vision had quicker times to first fixations in general, however on directions page users who did not wear glasses or contacts had faster times to first fixations. Users of glasses had faster first fixation times than contact users on the pages with jpeg images. Table 4. Correlation by Eyewear shows the correlations of validity by eyewear

are not significant, however the eyewear is significant when compared to actual data collected. This is an indicator that users who wore no vision correction devices had the longest fixations. However, the opposite order occurred in the number of fixations per minute, calculated by taking the number of fixations in a session, dividing it by the time it took to complete the session.

Table 4. Correlation by Eyewear

	Eyewear	Contacts	None
Left Eye Validity	-0.009	0.118	-0.068
Right Eye Validity	-0.134	0.066	-0.025
Average Fixation Duration	(**)0.641	(**)-0.661	(**)0.465
Average Fixation Per Minute	-0.286	(**)0.525	-0.299

The results in Table 5. Average Fixations by Eyewear show this inverse relationship between average fixation duration and the average fixations per minute count. For example, users of contacts may be perceived to have the most activity due to the fact they have the most fixations, but this is not the case, due to the fact the average duration is substantially less. One hypothesized reason for the breaks in fixations is that it is due to substantial periods of invalid data, caused by glare, reflections, or other light refractions coming off of the lenses of the contacts and/or glasses.

Table 5. Average Fixations by Eyewear

Eyewear	Average Fixation Duration (std)	Average Fixations per minute (std)
Contacts	514.91 (52.95)	174.84 (12.83)
Glasses	661.68 (65.60)	144.90 (26.75)
None	687.05 (100.18)	145.54 (22.37)
All Eyes	639.65 (108.29)	152.42 (24.47)

4.3 Gender

The same analysis was performed on gender, to find no significant findings in terms of pupil size, validity, fixation data, time on task, or task navigational movements.

4.4 Language

There was not enough data to determine significant results regarding language findings, but there are some indications that future research could be warranted. For example, participants with English as their native language took significantly less time to view the images than English-as-a-second-language participants. The average time for English-as-a-second-language participants was 5.98 minutes, where as participants who spoke English as their native language spent only 3.87 minutes on the study. In addition, on 4 AOI fixations both on direction pages and image pages, non-English

students had minor significance correlations of 0.408 to 0.840, indicating they spent more time gazing at AOIs, both in terms of number of fixations and fixation lengths. However, these findings were somewhat inconclusive due to a lack of more AOIs with significant or higher correlations. In addition, there was a lack of diversity in the first language of participants, limiting the conclusiveness of the findings even more. However, these preliminary findings will help in preparing future studies on these topics.

5 Summary and Implications for Future Research

In designing usability testing of transportation signs, studies should adjust fixation results for users based on type of eyewear or eye corrective treatment. Specifically, when testing between subjects, the demographics of both sets of subjects may need to have the same eyewear and eye color to yield the most accurate results. Otherwise, sign XYZ could be deemed less effective if testing with users wearing mostly contacts verses sign ABC, which was testing in a group of mostly non-corrective wear users. Hazel eyes and contacts gather the least valid eye tracking data, however there are no other significant differences found between the other eye colors or eyewear.

Implications for international design of usability studies include accounting for eye demographics, which may affect the sensors, and which may vary greatly from one country or locality to the next. Pupil size is of concern, as this was the most significantly correlated item to eye tracker's acceptance of a piece of data. In addition, eye color and eyewear have significant impacts, and can vary greatly from one country to the next. For example, lighter colored iris are most commonly found in Europeans and individuals of European admixture while darker iris colors are more common in the Middle Eastern and Southern Asian populations [7]. Also under consideration is the first language of the user, which can also vary from one part of the world to the next.

One research limitation is that we did not track users who may have had laser eye surgery. In addition, the affect of hazel eyes may be impacting the findings for contact lenses, and vice versa. There was a correlation of only 0.136 between eye color and eyewear, but brown eyes were significantly correlated to eyewear (-.481) and contacts (.484). While no other eye color or eyewear showed a significant relationship, a study with larger, more evenly distributed sample sizes may be needed to remove any impact one factor may have against another.

Future research areas include testing to see if eye validity errors are caused by (rather than just correlating with) eye color and eyewear and if so by how much, as well as to see if the invalid data is the true cause of making one fixation look like multiple fixations may be necessary to validate the need for usability testing where eyewear is controlled. In addition, testing needs to be done to see if this would happen on other eye tracking devices. More research is also needed to find what other fields this is affecting, and if there is a place to report false readings.

In addition, running a study that increased the sample size of the different language groups should provide more data on times to first fixation, fixation durations, etc, and perhaps with other physiological sensors to see if the user reacts in the intended way to the content read. English users take less observation time for English text, although further research is needed into languages that use an alphabet system that is visually

more complex. Complex alphabets require more strokes to construct the individually letter forms than simpler alphabets. For example, when comparing Latin and Hindi alphabets, Hindi would be more complex. Future research would look into the time it takes to read simple verses complex alphabets, to see if the adage, "we read best is what we read most true," is really true.

Lastly, we are not sure how the alphabet form/structure effects a user's observation time, verses the orientation of the letters. This is because some of the jpegs had text on an angle and the information was not organized off of a straight horizontal line in neat columns. This could have huge implications when designing messages in a text image, from transportation system signs to computer media images.

Acknowledgments. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. CNS-0619379. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Dr. Andrew Sears provided appreciated guidance and support in this project.

References

1. http://tobii.com/archive/pages/17744/view.aspx
2. http://www.eyegaze.com/content/eyetracking-research-tools#reliability
3. Nini, P.: Typography and the Aging Eye: Typeface Legibility for Older Viewers with Vision Problems. January 23, 2006. AIGI online Design Archives (2006)
4. Miniotas, D., Špakov, O., MacKenzie, I.S.: Eye gaze interaction with expanding targets. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, April 24-29, 2004, pp. 1255–1258. ACM, New York (2004)
5. Duchowski, A.T.: Eye Tracking Methodology Theory and Practice, 2nd edn. Springer, London (2007)
6. Wang, H., Lin, S., Liu, X., Kang, S.B.: Separating Reflections in Human Iris Images for Illumination Estimation. In: Proc. IEEE International Conference on Computer Vision (2005)
7. Sturm, R.A., Frudakis, T.N.: Eye colour: portals into pigmentation genes and ancestry. Trends in Genetics, vol.20.8 (2004)