

EXPLORING EYE TRACKING IN DESIGN EVALUATION

Sami Kukkonen, University of Art and Design Helsinki, Finland

Abstract

This paper describes one test from a larger eye tracking and design evaluation research aiming at developing methods for studying the perception of design products. The test focuses on searching for a connection between gaze, product attitude and preference when comparing the products in pairs and when selecting a favorite from the available products. The role of appearance and apparent usability in product attitude and preference are also explored.

The test results suggest that a correlation between gaze and product attitude can be found with certain indicators. Product preference can also be seen quite clearly in the visualized gaze data. Some distinct differences were found between the groups of designers and other test persons in the emphasis of appearance and apparent usability as variables influencing on both the product attitude and preference.

Keywords:

Design Evaluation, Visual Perception, Eye Tracking, Mobile Phone

Introduction

This paper presents an eye tracking test conducted in order to develop methods for studying design products with eye tracking. The test aimed to explore possible connections between people's gaze, product attitude and preference in comparing and evaluating the design products. At first the test settings and procedures are explained and then the various statistical and visual analysis approaches and their results are described. The discussion includes the lessons learned and the possible ways to improve the test and analysis methods. The paper is concluded with visions of the benefits and the means to utilize this research in the future.

The study was performed as part of a "Perception of Design" (-p-o-d-) research project, which is a co-operation of the University of Art and Design Helsinki and the University of Tampere and is funded by the National Technology Agency of Finland (TEKES) and the Academy of Finland. This project aims at developing methods for studying the perception of design products by combining eye tracking, gaze path analysis and design evaluation, and by utilizing these methods to compare people with and without design education and their perception of different design representations like sketches, photos and 3-d models.

Background

Most companies apply a range of metrics to evaluate the products they make, both during and after the development stage, in order to make informed decisions. Various methods exist for evaluating product design through usability, sales, profit, ease of manufacturing, and so on. The influences of various product attributes affecting to a product's general appeal have been examined widely, but the relevance of the products' visual properties has not been analyzed separately (Keinonen 1998). There seems to be a lack of suitable methods for measuring the appearance, the visible design, and its effect on customers' product preference. Therefore, the design decisions are often founded purely on subjective opinions.

Since visual evaluation depends directly on the sense of sight, tracking the gaze on the perceived target should provide measurable data for design evaluation. Eye tracking experiments have been performed for several decades already and on various research fields (Duchowski 2002). There are several earlier studies dealing with perception of pictures but they seldom involve design evaluation. Hammer (1992) surveyed the potentials and limits of eye tracking in examination of design products in his studies. One of the common findings with earlier picture viewing tests (e.g. Buswell 1935, Yarbus 1967) was that the given task or viewing motivation affects the gaze. This finding was also verified in our initial pilot test (Koivunen et.al 2004). Hammer and Lengyel (1989) concluded further that people are looking at what they feel most informative for the task. The study described in this paper attempts to explore deeper into the role of product design in the viewing motivation.

The two main attributes, appearance and apparent usability, were chosen under examination because they are intrinsic attributes which lie within a designer's direct influence (Keinonen 1998). They are clear physical and visual attributes which a designer can work on in order to achieve the desired impression. Apparent usability refers to perceived and expected usability as opposed to the real experience of using the product (Tractinsky 1997).

Eagly and Chaiken (1993) define attitude as a general and enduring positive or negative feeling about some person, object or issue. The preference implies here an action of making a choice based on personal views. The product attitude and preference may, but do not necessarily always correlate. Therefore they are considered as separate factors in this study.

Aims

The main goal of the test was to experiment with new ideas which could be applied to the combined eye tracking and design evaluation methods. The general research questions were:

- Does the gaze reveal anything about people's attitudes towards the compared products or about their product preference within the available products?
- Do the products' appearance and apparent usability play a role in the product attitude and preference for people with and without design background?

The more detailed sub-questions included issues like:

- Which product details draw the attention in comparison and evaluation situations?
- Will these details have any connection with the positive and negative details indicated by the testees in the post questionnaire?
- Do the left-right layout and the western reading direction affect the gaze in pair comparison screens?
- Do the pair comparison, the selection of the favorite and the post questionnaire evaluation produce comparable answers?

Test method

Apparatus

The test was performed with a Tobii 1750 eye tracker. The IR transmitters and the eye tracking camera are built inside a 17" TFT monitor so there was no need to attach anything to the test persons. Tobii tracks both eyes with 50 hz frequency. The gaze data was visualized as heat maps

with Tobii's ClearView software and then compiled with PhotoShop. The raw numerical data was exported to Excel where the statistical data was calculated.

Participants

There were 32 participants in the test of whom 4 were disregarded because of bad eye tracking data. Half of the remaining 28 people, 6 females and 8 males, had a design education background. The other half with no design background had an equal gender distribution. Age of the testees varied from 21 to 45 years with mean value of 31 years.

Material

Five mobile phones not marketed in Finland were selected for the test and their brand names were erased from the product pictures in Photoshop. The phone pictures were arranged in pairs so that each phone was paired once with all the other phones. The complete set of 10 screens presented each phone on four occasions and equally often on both left and right side of the screen. The last screen presented all the five phones side by side. All screens were 1280 by 1024 pixels and the phones appeared on the monitor roughly in 1:1 size.

Procedure

The idea of the test was to make people compare pairs of mobile phones and to divide ten points between the phones in each pair according to which one they liked better.

The test began with a calibration sequence and after everything was ready, the test sequence was explained and practiced with screen examples while the tracker was already recording.

Each comparison screen was shown for 8 seconds, after which a cartoon balloon with two question marks appeared below the phones to remind the testees to announce the points for each phone. A screen with a small colorful circle at the top edge was shown in between each comparison screen for one second to draw the subject's attention away from the area where the next phones would appear. These intervening screens were also used for verifying the validity of the gaze data visually. In the end all five phones were shown together and people were asked to pick their favorite.

After the test each testee filled a questionnaire about their backgrounds including education, profession, age, gender and possible problems with eyesight. The phone pictures were also evaluated more thoroughly. People were asked to mark positive and negative details in each phone and to give scores from 1 to 10 to each phone's appearance and apparent usability. Similar scores were given to the personal importance of five possible factors effecting in the choice of a mobile phone; brand, appearance, price, usability and technical features. Finally people's familiarity with the phones was inquired, even though it was assumed that they had not previously seen them.

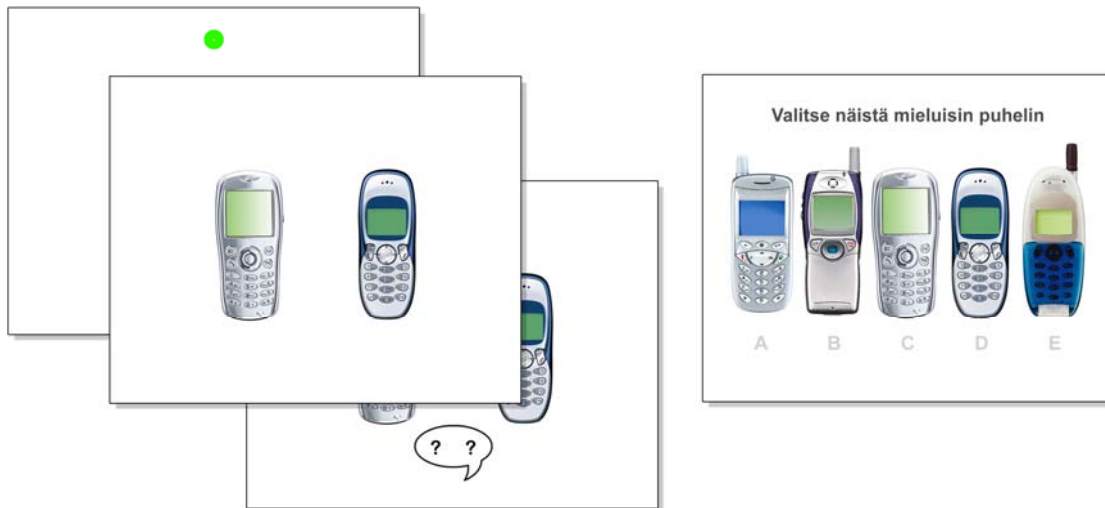


FIGURE 1. Samples of the pair comparison and the favorite selection screens.

Analysis

Parameters

There are several parameters which can be considered useful in statistical analysis of the gaze data. These parameters are based on fixations and saccades, and their durations and locations at the stimuli. Fixations mean the moments when the eyes are relatively still and focused on a certain target. The common average fixation duration is between 200 and 300ms. The saccades are rapid movements between the fixations, during which the brain does not receive visual information from the eyes.

Jacob and Karn listed six most commonly used metrics from 24 reviewed usability research eye tracking studies: overall fixation count, percentage of the total time spent on each area of interest, average fixation duration, fixation count on each area of interest, average dwell time on each area of interest, and overall fixation rate. Although these were the most commonly used metrics, some of them are not always considered most useful. For instance, the overall fixation count should be related to the time spent on the task rather than used alone. (Jacob and Karn 2003).

There are several definitions of fixations and their duration, and various algorithms are used for calculating them. With the default settings of the Tobii's ClearView software, the fixation size was set at 50 pixels, which corresponds to the foveal zone, the small area of high acuity in human vision. The minimum fixation duration was set to 70ms.

Arranging and analyzing the data

In the analysis phase the raw gaze data was arranged and cut to sections according to the viewed screens. Only the automatically timed comparison screens were used in numerical analysis, since the time each person took to announce their answers during the following screen varied. All the comparison and evaluation points, the selection answers, post questionnaire evaluations and backgrounds were also compiled to enable comparisons with the statistics and visualized heat maps produced from the gaze data.

The idea was to compare the gaze data with the given comparison scores, the selected favorite phones and the individual phone evaluations from questionnaires. The fixation counts, durations

and their averages were calculated per phone and per person, and these were compared with the given comparison points. The first fixations on and the transitions between the areas of interest were also defined and counted. These were then calculated according to different phones and also according to layout's left and right side.

Cumulative gaze duration heat maps of each phone were compiled from individual pair comparison heat map screens in the PhotoShop for visual analysis. The negative and positive areas marked in the questionnaires were then applied to these heat maps and to the favorite selection screen heat maps.

Results

Layout and viewing order

The results from analyzing the first fixations in each comparison screen suggest that the western reading direction may play a role in viewing and comparing the pictures in a left-right layout. The left picture was looked at before the right picture in 78% of the screens. Seven people (1 designer, 6 others) out of the 28 testees looked constantly first at the left picture. Becoming familiar with the phone pictures seemed to decrease the dominance of the left side towards the end of the end of the 10 screens though.

The left picture also had a higher fixation count average, (12.1 fixations per screen) than the right picture (10.1) and the average total viewing time spent on the left phone was also higher than on the right one (2796ms vs. 2283ms).

The way of viewing seemed to favor the left phones, but overall the left and right sides still received equal amount of comparison points (1400 vs. 1400) and the phones on both sides were liked better than the opposites almost equally often (117 vs. 113 times, in 50 cases phones received equal points).

Attention, average fixation counts and durations per phone

The average fixation count on each phone correlated well (0.92) with the given comparison point totals. However, it did not exactly match the phones' point rank order because the second best ranked phone had a slightly higher fixation count average.

An interesting difference was found between the groups of designers and others. Designers' average fixation count correlated well with the appearance evaluation scores (0.90) in the questionnaire, but not at all with the evaluation of apparent usability (-0.57). For the group of others appearance (0.70) correlated worse than apparent usability (0.84). This might suggest differences in evaluation interests and motivations.

The average fixation durations correlated with the comparison points worse (0.81) than the average fixation counts. The designers' correlation (0.85) was fairly high, but the others' only very low (0.36).

Questionnaire evaluation, comparison points and selecting the favorite phone

Similar difference between designers and others was found in connection between the comparison points given during the test and the phone evaluation scores in the questionnaire. Designers may have based their comparison and personal preference much more on appearance than on apparent usability, whereas the others have considered both attributes.

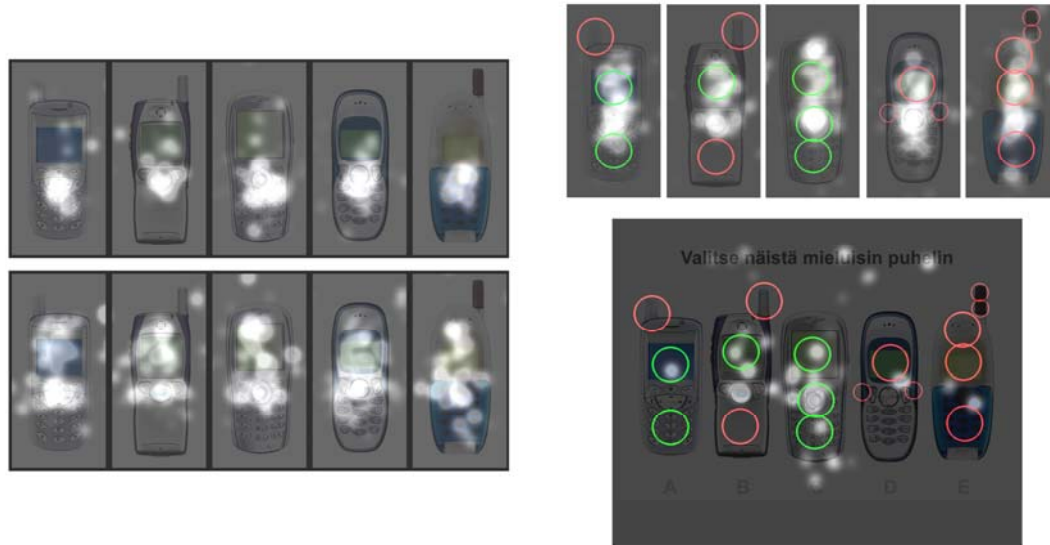


FIGURE 2. Left: Cumulative heat maps from pair comparison presenting two people with opposite (narrow and wider) eye movement behavior. Right: Sample of one person's cumulative pair comparison heat maps and a selection screen heat map with positive (green) and negative (red) details indicated in the post questionnaire.

Discussion

The original idea of defining the individual product details viewed during the pair comparison proved to be impossible. The desire to keep the phone pictures close to real life size made the details so small that one could not be differentiated adequately from another. For a study of the product details the stimuli should be presented as large as the screen permits. However, the general observations from some of the images were found parallel to the findings of Babcock et al. (2003) as the areas receiving most attention in the picture during recording were not always the same as the ones pointed out afterwards as decisive elements (Fig 2. right). These images also correlated with the wider and more narrow eye movement behaviors mentioned already in Buswell's (1935) early studies.

The selection screen heat maps often revealed people's product preference already in a superficial examination (Fig 2. bottom right). The overall average fixation counts seemed to correlate with the overall average order of the attitude towards the phones. Thus it seems that the more liked phones are viewed more intensely, but it is unknown, how much the inquiry of the positive product attitude might affect the gaze. A reversed task, focusing on worse phones, could have produced comparable answers, but possibly different kind of gaze data. In the future, in the real design evaluation cases, much more attention should be paid to formulating the questions presented to the test participants, and to using control groups with altered stimuli order and possibly reversed questions to verify the data.

The product attitude, the product preference and the evaluation ratings do not always seem to correlate. This might be caused by some people comparing and evaluating the phones in a more technical and objective way while considering their personal reasons only when selecting the

favorite. However, both of the above mentioned test aspects could be useful in design evaluation or in revealing the background motivations of different groups of people.

Vuori et al. (2004) found saccade duration a useful parameter in their recent image quality study. The same was attempted, but found useless in our study because of the eye tracker's 50hz frequency. Saccades are too fast (typically less than 50ms) to be recorded in only 20ms intervals.

Since this test was part of a larger test series where several method development related issues were tried out all the different test sections were recorded in a row. This produced close to 30 minutes of eye tracking data from each participant and caused lots of extra work later in cutting the data into proper packages and, therefore, making the whole analysis process generally slower. Even if the test sessions became more complicated when recorded in smaller consecutive sections, such action would save much time later in the analysis phases.

Conclusions

The findings in this study show that it is possible to find correlations between gaze and different aspects of design evaluation. These findings may prove useful in developing the combined eye tracking and design evaluation methods further. When applied to real design cases these methods could assist a designer in various stages where the visual aspects of the product designs need to be evaluated. Hammer and Lengyel (1991) summarized how eye tracking can help a designer: "It is undoubtedly useful for measuring attention, to determine the most attractive areas of a product." With given tasks, it "can indicate which product elements carry the given meanings or brand identification". The eye tracking was not found useful though, when the meanings are carried by the whole of a product. Since their studies the eye tracking technology has advanced and the computing power has increased considerably. Therefore, more accurate and detailed eye tracking studies could be performed in the challenging field of design evaluation, at present and in the future.

One possible future application could involve comparing the intended product claims to the messages of the product's appearance. The results could be utilized to define areas and details requiring special attention in a design facelift. Another facelift approach could investigate which product details carry the brand identity and how well the existing products stand out from the mass of similar products.

These same approaches could also be applied to studying new visualized design concepts. Different tasks could be used to compare the attention received by the various product details and their intended messages, and to examine their role in product preference. A study of competitors' products or company's own older products could also provide valuable data for the basis of a new product concept.

These application scenarios are not aiming at replacing any part of designers' intuitive and creative work. On the contrary, they could enhance the evaluation of alternative designs and the communication within a development group with new views and fresh discussion.

Acknowledgments

I would like to thank the other members of our project group, Kimmo Koivunen, Sami Lahtinen, Harri Rantala and Selina Sharmin, for their co-operation in conducting the test and in arranging the raw gaze data, and Prof. Turkka Keinonen for his supervision and helpful comments for this paper.

References:

- Babcock, J., Pelz, J. & Fairchild, M. (2003). Eye Tracking Observers During Rank Order, Paired Comparison, and Graphical Rating Tasks. In Proceedings of the 2003 PICS Digital Photography Conference, Vol. 6. Rochester, NY, pp. 10-15.
- Buswell G. T. (1935). How People Look at Pictures, A Study of the Psychology of Perception in Art. The University of Chicago Press, Chicago.
- Duchowski, A.T. (2002). A Breadth-First Survey of Eye Tracking Applications. In Behavior Research Methods, Instruments & Computers (BRMIC) 2002 34(4). pp. 455-470.
- Eagly, A. & Chaiken, S. (1993). The Psychology of Attitudes. Harcourt College Publishers, New York.
- Hammer, N. and Lengyel, S. (1989). Do people perceive what the design expresses? In Vihma, S. (ed.). Semantic Visions in Design, the Symposium on Design Research and Semiotics. UIAH, Helsinki.
- Hammer, N. & Lengyel, S. (1991). Identifying Semantic Markers in Design Products: The Use of Eye-Movement Recordings in Industrial Design. In Schmid, R. & Zambarbieri, D. (eds.). Oculomotor Control and Cognitive Processes - Normal and Pathological Aspects. Elsevier Science, Amsterdam, pp. 445-455.
- Hammer, N. (1992). Möglichkeiten und Grenzen der Überprüfung von Designprodukten durch Okulometrie. Die Blaue Eule, Essen.
- Jacob, R. and Karn, K. (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. (Section Commentary) In Hyona, J., Radach, R. & Deubel, H. (eds.). The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research. Elsevier Science, Oxford.
- Keinonen, T. (1998). One-dimensional usability: Influence of usability on consumers' product preference. Taideteollinen korkeakoulu, Helsinki, pp. 64-90 and 111-138.
- Koivunen, K., Kukkonen, S., Lahtinen, S., Rantala, H. and Sharmin S. (2004). Towards Deeper Understanding of How People Perceive Design in Products. In Proceedings of the Computer in Art and Design Education (CADE), Copenhagen.
- Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In CHI'97 Conference proceedings, Atlanta, 22-27 March. ACM, New York, pp. 115-122.
- Vuori, T., Olkkonen, M., Pölonen, M., Siren, A. & Häkkinen, J. (2004). Can Eye Movements Be Quantitatively Applied to Image Quality Studies? In NordiCHI '04 Conference proceedings. ACM Press, pp. 335-338.
- Yarbus, A. L. (1967). Eye Movements in Vision. Plenum Press, New York.