Intrusive and Non-intrusive Evaluation of Ambient Displays

Xiaobin Shen
xrshen@unimelb.edu.au
University of Melbourne

Peter Eades
peter.eades@nicta.com.au
National ICT Australia
University of Sydney

Seokhee Hong
seokhee.hong@nicta.com.au
National ICT Australia
University of Sydney

Andrew Vande Moere
andrew@arch.usyd.edu.au
University of Sydney

ABSTRACT

This paper addresses two problems: “What are the appropriate methods for evaluating information systems?” and “How do we measure the impact of ambient information systems?” Inspired by concepts in the social and behavioral science, we categorize the evaluation of ambient displays into two styles: *intrusive* and *non-intrusive*. Furthermore, two case studies are used to illustrate these two evaluation styles. An intrusive evaluation of *MoneyColor* shows that the correct disruptive order for ambient displays is animation, color, area and shape. A non-intrusive evaluation of *Fisherman* proposes an effectiveness measurement, and reveals three issues to improve the effectiveness of ambient displays.

Keywords

Ambient displays, intrusive evaluation, information visualization, human computer interaction

INTRODUCTION

Ambient displays to some extent come from the ubiquitous computing dream, which was first proposed by Weiser [1]. Following his dream, many pioneers of ubiquitous computing have created a plethora of overlapping terminology (for example, disappearing computing [2], tangible computing [3], pervasive computing [4], peripheral display [5], ambient display [6], informative art [7], notification system [8], or even ambient information system [9]). The differences between some of these terms are not obvious. In this paper we use the term “ambient displays” generically.

Research on ambient displays is still immature, and there is no universally accepted definition available. Ishii et al. [3], Matthews et al. [5], Stasko et al. [9] and Mankoff et al. [6] all propose their own definitions. Here we follow Stasko [9]: “ambient displays typically communicate just one, or perhaps a few at the most, pieces of information and the aesthetics and visual appeal of the display are often paramount”.

Many ambient displays have been designed and developed but less progress has been made in the evaluation of these displays. However, good evaluation methods can judge the quality of the design to provide a basis for making improvements, and we believe that evaluation methods should be a priority for researchers.

This paper focuses on two issues: “What are the appropriate methods for evaluating information systems?” and “How do we measure the impact of ambient information systems?” These questions are difficult and will take some years of effort to settle. In this paper we propose a concept that may play a role in the answers. More specifically, two evaluation styles, intrusive and non-intrusive evaluation, are proposed in the next section. Following this, we illustrate the styles with two case studies.

INTRUSIVE AND NON-INTRUSIVE EVALUATION

Many researchers have realized the importance of the evaluation of ambient displays. Mankoff et al. [6] proposed a heuristic evaluation for ambient displays. Pousman et al. [9] proposed a four design dimension to guide in the evaluation of ambient information systems. McCrickard [8] proposed an IRC framework to evaluate the notification system. Shami [10] et al. proposed the CUEPD evaluation method to capture context of use through individualized scenario building, enactment and reflection.

McGrath [11] categorized eight normal evaluation methods in the social and behavioral science and classified them by two dimensions: “Obtrusive vs. Unobtrusive” and “Abstract vs. Concrete” (See Figure 1).

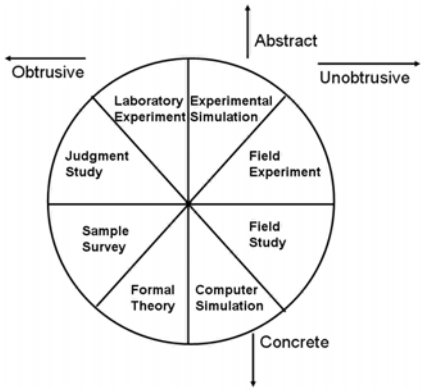

Figure 1. Evaluation Methods and Classification

These eight evaluation methods are applied broadly in behavioral and social science, but can be used in information visualization and even ambient displays.

Inspired by McGrath's classification, we proposed two new terms "intrusive" and "non-intrusive" for ambient display evaluation.

Intrusive Evaluation — is where the user's normal behavior is consciously disrupted by the evaluation experiment. This kind of evaluation often consists of usability tests in a laboratory environment for a short period. Most such experiments are conducted using well established evaluation techniques in information visualization (for example, questionnaires and interviews).

Non-Intrusive Evaluation — is where the user's normal behavior is not consciously disrupted by the evaluation experiment. This often focuses on actual use in a general environment (*in situ*) over a long period. Currently, few existing evaluation techniques can be applied successfully in this manner.

Intrusive and non-intrusive seem more like endpoints on a continuous range than buckets for evaluation methods. The difference between these two evaluations is the level of user involvement. Intrusive evaluation seems to be good at quantitative measurement of parameters, but non-intrusive evaluation may be not. On the other hand, intrusive evaluation may have higher cognitive load, which leads to affect the validity of results, but non-intrusive evaluation can have better results by having lower cognitive load.

Two case studies are described in the next section to illustrate these two evaluation styles.

TWO CASE STUDIES

In this section, we describe evaluations of two systems: *MoneyColor* and *Fisherman*. Both systems represent real-time information as well as decorating the architectural space. Both are designed for the public sites.

The data used in *MoneyColor* is stock price and volume from the Australian Stock Exchange. More specifically, our experiment used BHP-Billiton[1] price and volume data.

The *MoneyColor* display is inspired by the art of Hans Heysen [12] an Australian watercolor painter of the early 20th century. Paintings in the style of Heysen form a peaceful background, and are often used simply as decoration in Australia, from homes to boardrooms.

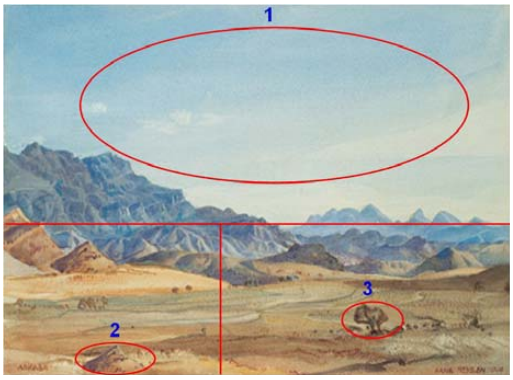

Figure 2. Metaphor of *MoneyColor*

There are three metaphors in *MoneyColor* (see Figure 2):

1. The color of the sky represents the general stock index. The darker the sky, the lower the general stock index.
2. The position of a specific mountain represents the current BHP stock price. The higher the position of the mountain on the image, the higher the stock price.
3. The size of the tree represents stock volume. The larger the tree, the greater the stock volume.

MoneyColor is aimed for use by stock holders and brokers.

The data source used in *Fisherman* is statistics. More specifically, we use three parameters of the NICTA[2] website: the number of hits on the web page; the bandwidth of the web server, and the number of pages viewed.

There are three metaphors in *Fisherman* (see Figure 3):

1. The level of fog in the mountain represents the number of hits on the web page. Heavier fog indicates fewer hits.
2. The number of trees represents the number of viewed pages. The more trees, the more pages viewed.
3. The position of the boat represents the bandwidth of the server. The higher the position, the higher the bandwidth.

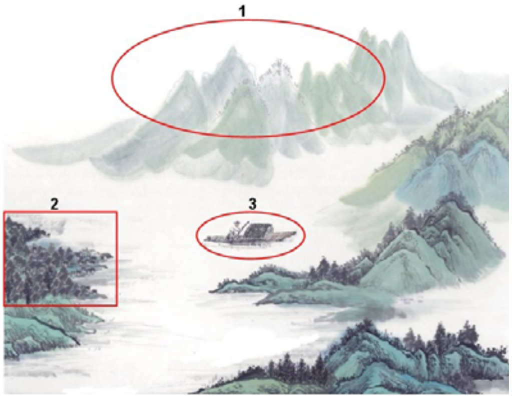

Figure 3. Metaphor of *Fisherman*

[1] BHP-Billiton is a large mining company based in Australia; price movements in BHP-Billiton are very influential on the Australian Stock Exchange.

[2] National ICT Australia, an Australian Government research laboratory.

Intrusive Evaluation of *MoneyColor*

The general evaluation aim of *MoneyColor* is to explore the way in which different factors disrupt the user. More specifically, we want to determine the order of disruptiveness of the following factors:

- *Animation*: the image morphing technique
- *Color*: the change in hue.
- *Area*: the change in the size of the tree
- *Position*: the change of the location of the mountain.

We hypothesize that the correct disruptive order for ambient displays is: animation, color, position, then area. This hypothesis is loosely based on the results of Cleveland and McGill [13] on the order of visual cues for effectiveness in graphical presentation.

The experiment was conducted in a visualization laboratory. Eighteen (nine female) subjects participated in this experiment. Subjects ranged from 21 to 35 years (10 masters, 6 PhD and 2 Post-doc). Seven subjects knew nothing about ambient displays and the remainder had some knowledge; none were experts.

The experiment use *Square-Click,* a simple game that dynamically assigns a random location for a black square (size 80*80 pixels) every second. Subjects need to mouse-click the black square within one second of its appearance (see Figure 4). If successful, the black square will be assigned to a new random location and the user scores 1. If not, the black square will be assigned a new location after one second and the user scores 0.

The experiment used a standard PC with two standard 19 inch monitors with a resolution of 1024*768, and one LogiTech QuickCam Pro4000 web camera together with face detection software. A mouse was the only user input device. The monitors, camera, chair, and desk were arranged as in Figure 5.

There were two user tasks in this experiment. The primary task was to play *Square-Click*; this ran for two minutes on the "focus" monitor, after which the user score was recorded. The secondary task was to obtain information about BHP stock via *MoneyColor* on the "peripheral" monitor. Participants were encouraged to not only get a good score in the primary task but also get BHP stock information.

The experiment included fifteen tests. Each test had *Square Click* system plus *MoneyColor*, but focused on different factors:

- Test 2-3: "color" with/without animation;
- Test 4-5: "position" with/without animation;
- Test 6-7: "area" with/without animation;
- Test 8-9: "color and position" with/without animation;
- Test 10-11: "color and area" with/without animation;
- Test 12-13: "position and area" with/without animation;
- Test 14-15: "color, position and area" with/without animation.

Each test lasted two minutes and was followed by a two-to-four minute break, so the entire experiment lasted around one and a half hours. Testing of all 18 subjects was conducted within three weeks.

A questionnaire was also used at the end of each test to collect additional information, with three Likert-scale questions:

- Does the value of the parameter change?
- How does the value of the parameter change?
- How much does the value of the parameter change?

Further, the face detection software is used to record whether subjects look at *MoneyColor* or not.

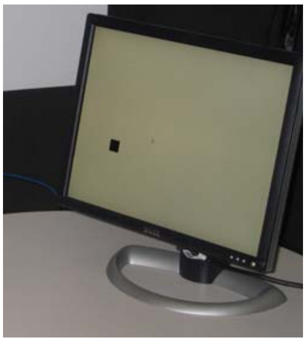

Figure 4. Square Click

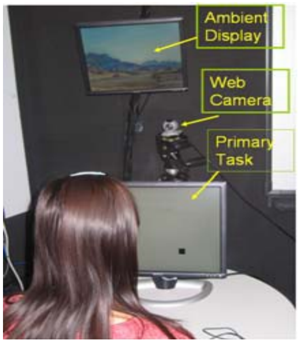

Figure 5. Actual Settings

Non-Intrusive Evaluation of *Fisherman*

The general aim for this experiment is to discover the relationship between *comprehension* and time. We hypothesize that the comprehension of subjects to *Fisherman* increases with time.

This experiment was conducted in a public corridor opposite to an elevator, close to a public facilities room. The display was in a purpose-built frame, which also enclosed an IR sensor and a camera[3]. Furthermore, the *Fisherman* metaphor was described on an A4 size paper on the new frame (see Figure 6).

Every person passing by *Fisherman* was a subject of this experiment. These people are mainly researchers. Most have some knowledge of ambient displays but none is an expert. The whole experiment lasted six months.

Subjects were randomly chosen to fill the questionnaire and subjects were not allowed to look at the display during answering questions. Three questionnaires were scheduled within the six months. Each questionnaire was mainly to

[3] Since the system included a sensor and camera in a semi-public place, legal opinion was obtained to ensure that the system complied with privacy legislation.

measure three attributes: *comprehension*, *usefulness* and *aesthetics*.

The comprehension questions were:

CQ1: Does *Fisherman* convey any information?

CQ2: How many types of information are represented in *Fisherman*?

CQ3: What kind of information does *Fisherman* represent?

CQ4: Have you ever noticed changes in *Fisherman*?

The usefulness questions were:

UQ1: Is *Fisherman* useful to you?

UQ2: Why?

The aesthetic questions were:

AQ1: Do you think *Fisherman* is visually appealing?

AQ2: If possible, would you put *Fisherman* in your home/office?

AQ3: Why?

The IR sensor and camera recorded the number of people passing the display and the number of people who turned their face toward the display.

There was no specific primary task designed for the subject; almost all the subjects were engaged in a normal everyday primary task such as using the elevator or the facilities room. Subjects shifted focus to the display to obtain information; this was a secondary task.

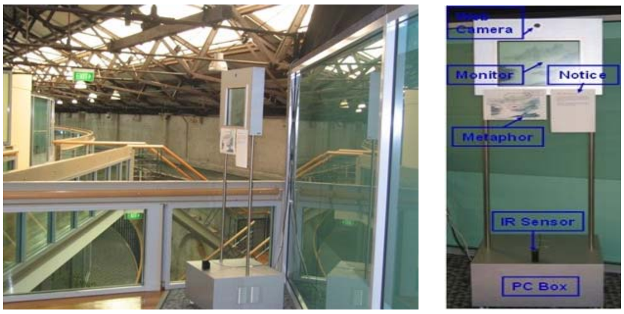

Figure 6. Implementation of Fisherman

Results of the Intrusive Evaluation of *MoneyColor*

A within-subject experimental design was used with non-fixed ordering of the experimental tests. Three parameters are analyzed:

- *Mean Comprehension Rate* (MCR) is derived from the answers given in the questionnaire; it measures the correctness of the information that subjects recalled about the information on the peripheral display. A larger MCR indicates better understanding of the ambient display.
- *Mean Self-Interruption* (MSI) counts the number of focus shifts to the peripheral screen prompted by the subjects themselves; a larger MSI denotes a more curious or nervous subject.
- *Mean Display-Distraction* (MDD) counts the number of focus shifts to the peripheral screen caused by display distraction; a lower MDD denotes a calmer ambient display.

The difference between Mean Self-Interruption (MSI) and Mean Display-Distraction (MDD) is not subtle. As a gross simplification, we assume any glance after a display update contributes to the Mean Self-Interruption (MSI).

Two major results are discussed below.

MCR	**Test 2-3**	**Test 4-5**	**Test 6-7**	**Test 8-9**	**Test 10-11**	**Test 12-13**	**Test 14-15**
Animation	0.84	0.67	0.83	0.80	0.60	0.54	0.53
Static	0.83	0.65	0.80	0.74	0.58	0.58	0.52
p	0.05	0.04	0.04	0.03	0.05	0.05	0.05

Table 1. Mean Comprehension Rate in Each Test

Table 1 shows that the value of Mean Comprehension Rate (MCR) with animation is higher than without. Furthermore, this difference is significant ($p<0.05$).

Also, Table 1 reveals that color (Test2-3) has the highest Mean Comprehension Rate (MCR) and position (Test4-5) achieves the lowest (as a single factor).

	One Visual Cue	**Two Visual Cues**	**Three Visual Cues**
MCR	0.771	0.633	0.531
p	0.123	0.081	0.069

Table 2. Relationship between MCR and visual cues

Table 2 shows that the value of Mean Comprehension Rate (MCR) decreases with the increase in the number of visual cues. However, this result is not significant and requires further study.

A statistics correlation method was used to calculate relationships between Mean Comprehension Rate (MCR), Mean Self-Interruption (MSI) and Mean Display-Distraction (MDD) and results showed that Mean Comprehension Rate (MCR) is directly proportional to Mean Display-Distraction (MDD). However, there is no obvious relationship between Mean Comprehension Rate (MCR) and Mean Self-Interruption (MSI) (see Figure 7).

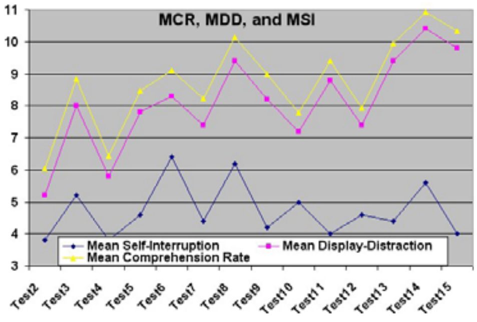

Figure 7. Relationship between MCR, MDD and MSI

Results of Non-Intrusive Evaluation of *Fisherman*

A standard deviation (STD) statistical method is used to analyze results. Three parameters are analyzed in the non-intrusive evaluation of *Fisherman.*

- The *Mean Comprehension Rate* (MCR), based on the answers from the comprehension questionnaire (CQ1-CQ4). A larger MCR indicates better understanding of the display.
- The *Total number of Subjects Passing by* (TSP) *Fisherman* in one day, measured using the IR sensor.
- The *Total number of Subjects Looking at* (TSL) *Fisherman* in one day, measured by the facial detection system.

It is clear that TSL ≤ TSP, but TSP also counts subjects passing by *Fisherman* without looking at the display. Thus we propose an effectiveness measurement ES as:

$$ES=TSL/TSP$$

Two major results on effectiveness are discussed below.

	1st MCR	2 MCR	3 MCR
CQ1	90.1%	100%	100%
CQ2	72.7%	76.9%	79.9%
CQ3	45.5%	76.9%	78.4%
CQ4	45.5%	69.2%	69.9%

Table 3. Results of Mean Comprehension Rate

Results from Table 3 show that Mean Comprehension Rate (MCR) in each question increases with time. This result supports our hypothesis that comprehension of *Fisherman* increases over time.

Table 4 shows the mean effectiveness value with standard deviation in each week (the first value in each cell is the mean effectiveness value; the second value is the standard deviation). From Table 4, it seems that effectiveness decreases over time.

	Week 1	Week 2	Week 3	Week 4
Sep., 05	34.8%/0.1	32.9%/0.2	16.9%/0.12	16.7%/0.1
Oct., 05	8.4%/0.03	9.0%/0.04	8.1%/0.03	7.2%/0.01
Nov, 05	7.4%/0.03	6.1%/0.02	5.7%/0.03	5.3%/0.02
Dec., 05	4.3%/0.02	4.7%/0.01	4.1%/0.01	Holiday
Jan., 05	Holiday	4.1%/0.01	3.9%/0.01	4.1%/0.01

Table 4. Mean value of effectiveness in each week

Discussion of Intrusive Evaluation of *MoneyColor*

The intrusive evaluation of *MoneyColor* shows that animation is the most disruptive factor. In fact, Table 1 shows that we can order the factors by disruptiveness as follows: Animation, Color, Area and Position. This result differs from the finding of Cleveland and McGill [13] (that the correct order for quantitative data is: position, area, color and animation). However, Cleveland and McGill investigated displays that demand full user attention, while we are investigating ambient displays. Thus this result shows a clear distinction between ambient and focal visualization.

Conclusion 1. The correct disruptive order for ambient displays is animation, color, area and shape.

Results in the *MoneyColor* evaluation (Figure 7) also show that more display distraction gives the better performance in comprehension. On the other hand, there is no obvious relationship between self-interruption and comprehension. Part of the reason is that display distraction is caused by the change of data source, whereas self-interruption depends on the personality of the subjects. Thus subjects have a better chance of identifying changes in *MoneyColor* by display-distraction than by self-interruption. This result is consistent with Matthews' finding on distraction (which she called "notification") [5]. Our result further adds weight to the hypothesis that different levels of display-distraction may affect the comprehension of ambient displays.

Conclusion 2. Better control of the level of display-distraction seems to enhance the level of comprehension for ambient displays.

Discussion of Non-Intrusive Evaluation of *Fisherman*

Results in the evaluation of *Fisherman* show that the effectiveness in *Fisherman* is quite low. Three reasons are reached to explain this:

1. *The data source does not interest users* — many subjects comment that the data source used in *Fisherman* was not related to their everyday activities. A typical comment: "I felt the display was interesting rather than directly useful, as the information represented here is not relevant to me. Visualizing statistical information of NICTA internet traffic has not affected my internet usage (it didn't bring any personal advantage to me). I'd like to see information about my activity that doesn't affect my privacy". This kind of comment implies the following conclusion:

Conclusion 3. Customization of data source can improve the comprehension of ambient displays.

2. *Lack of reference in the visual metaphor* — some subjects have difficulty interpreting information from the small changes in the metaphor used in *Fisherman*. A comment from one subject was: "I notice the color, the number of trees and the position of the boat changing but I can't get precise information from this change. Also I can't tell the difference between small percentages of change in these three metaphors. There is a lack of reference for the difference between heaviest and heavier fog." These comments imply:

Conclusion 4. Metaphors for quantitative measurements need some clues to be interpreted well.

3. *Subjects need a better way to interpret ambient displays* — ambient display is a new type of

visualization style and most subjects still prefer to access information by focal displays. A typical comment: "I only look at the display a couple of times a day and it seems to act as a cue for conveying information. But I still like normal visualization styles". This comment indicates that users need better support to interpret the information from ambient displays.

Conclusion 5. Users need better support information from ambient displays.

A significant question in many evaluations is: "When a test should be conducted?" Most researchers answer this question based on experience, but this evaluation attempts to use the pre-defined *effectiveness* measurement to determine the optimum time for the evaluation. Our case study has shown that the evaluation of *Fisherman* should be delayed until the value of *effectiveness* becomes stable. This is because a stable *effectiveness* value for *Fisherman* means that the display itself integrates into the environment and will not draw unusual attention from users: this meets the definition of non-intrusive evaluation of ambient displays.

Conclusion 6. Non-intrusive evaluation cannot be tested until the display integrates into the environment.

CONCLUSION

This paper focuses on discussing two questions:

1. "What are the appropriate methods for evaluating information systems?"
2. "How do we measure the impact of ambient information systems?"

To answer the first question, we present two evaluation styles: intrusive and non-intrusive evaluation. Two case studies are conducted by applying these two styles and six conclusions are draws from these two case studies.

Answers to the second question are mainly derived from the non-intrusive evaluation styles. We simply propose a quantitative effectiveness measurement to quantify the impact of *Fisherman*. As we believe the more subjects like the display, the better impact of the display.

This work is still in progress. Our future plans include more experiments to gain experience in the two evaluation styles. We aim to define the strengths and weaknesses of each style.

REFERENCES

1. Weiser, M. The computer for the 21st century. *Scientific American*, 1991. 265(3), 66-75
2. Disappearing Computer. Available at http://www.disappearing-computer.net
3. Ishii, H. et al. Tangible bits: towards seamless interfaces between people, bits and atoms, *in Proceedings of CHI'97* (Atlanta, USA), ACM Press, 234-241.
4. IBM Pervasive Computing. Available at http://wireless.ibm.com/pvc/cy/
5. Matthews, T., et al., A toolkit for managing user attention in peripheral displays, in *Proceedings of UIST'04* 247-256.
6. Mankoff, J., et al., Heuristic evaluation of ambient displays, in *Proceedings of CHI'03*, ACM Press, 169-176
7. Future Application Lab, Available at http://www.viktoria.se/fal/
8. McCrickard, D.S., et al., A model for notification systems evaluation--Assessing user goals for multitasking activity, *ACM Transactions on Computer-Human Interaction*, ACM Press, 10(4), 312-338.
9. Pousman, Z. et al., A taxonomy of ambient information systems: Four patterns of design. In *Proceedings of the AVI'06*, ACM Press 67-74.
10. Shami et al., Context of use evaluation of peripheral displays. In *Proceedings of the INTERACT'05*. Springer, 579-587.
11. McGrath J.E., Methodology matters: Doing research in the behavioral and social science, *Human-computer interaction: toward the year 2000*, Morgan Kaufmann Publisher, 152-169.
12. Heysen, H., Available at http://www.hawkersa.info/heysen.htm
13. Cleveland, W.S. et al., Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 1984. 79(387), 531-546.
14. Intel Open CV, Available at http://www.intel.com/research/m