remarks

• Grades Assignment 2 online.

• Exam (July 1, 11-13h):
  – 75% multiple choice
  – 25% open questions
  – Question(s) on guest lectures included.
  – We’ll try to have it graded the same day. But, don’t shoot is if we haven’t. ;)

• Today: Chapters 13 and 14

• Next Friday …
  – a very nice guest lecture!
  – deliver assignment 3 (before mid night)
Chapter 13
Introducing evaluation
The aims

• Key concepts and terms
• Types of evaluation methods.
• Evaluation methods for different purposes at different stages of the design process, and in different contexts of use.
• Mixing and modifying evaluation methods to meet the demands of evaluating novel systems.
• Challenges
• Illustrate how methods from Chapters 7 and 8 are used in evaluation.
Why, what, where and when

Iterative design and evaluation is a continuous process that examines:

- **Why**: to check users’ requirements and that they can use the product and they like it.
- **What**: a conceptual model, early prototypes of a new system and later, more complete prototypes.
- **Where**: in natural and laboratory settings.
- **When**: throughout design; finished products can be evaluated to collect information to inform new products.
Why you need to evaluate

“Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce successful results. If you don’t have user-testing as an integral part of your design process you are going to throw buckets of money down the drain.”

• source: http://www.asktog.com/columns/037TestOrElse.html
• Check the site for more topical discussions about design and evaluation.
Types of evaluation

• Controlled settings involving users, e.g., usability testing and experiments in laboratories and living labs (e.g., see Janssen et al., 2013 and Van den Broek, 2013).

• Natural settings involving users e.g., field studies and in the wild studies to see how the product is used in the real world (e.g., see Janssen, Van den Broek, & Westerink, 2012).

• Settings not involving users e.g., to predict, analyze, and model aspects of the interface analytics (e.g., see Van den Broek, Kisters, & Vuurpijl, 2005).
Living labs

- People’s use of technology in their everyday lives can be evaluated in living labs.

- Such evaluations are often context-sensitive and too difficult to do in a usability lab.

- At home or within cities: a network of sensors and audio/video recording devices.
Lab & field studies can compliment

- Field study to evaluate initial design ideas and get early feedback
- Make some design changes
- Usability test to check specific design features
- Field study to see what happens when used in natural environment
- Make some final design changes

**Figure 13.1** Example of the way laboratory-based usability testing and field studies can complement each other
Emotional and psychophysiological responses to tempo, mode, and percussiveness

Marjolein D. van der Zwaag
Philips Research Laboratories, The Netherlands

Joyce H.D.M. Westerink
Philips Research Laboratories, The Netherlands

Egon L. van den Broek
Human-centered Computing Consultancy, Austria; Human Media Interaction, University of Twente, The Netherlands

Abstract

People often listen to music to influence their emotional state. However, the specific musical characteristics which cause this process are not yet fully understood. We have investigated the influence of the musical characteristics of tempo, mode, and percussiveness on our emotions. In a quest towards ecologically valid results, 32 participants listened to 16 pop and 16 rock songs while conducting an office task. They rated experienced arousal, valence, and tension, while skin conductance and cardiovascular responses were recorded. An increase in tempo was found to lead to an increase in reported arousal and tension and a decrease in heart rate variability. More arousal was reported during minor than major mode songs. Level and frequency of skin conductance responses increased with an increase in percussiveness. Physiological responses revealed patterns that might not have been revealed by self-report. Interaction effects further suggest that musical characteristics interplay in modulating emotions. So, tempo, mode, and percussiveness indeed modulate our emotions and, consequently, can be used to direct emotions. Music presentation revealed subtle different results in a laboratory setting, where music was altered with breaks, from those in a more ecologically valid setting where continuous music was presented. All in all, this enhances our understanding of the influence of music on emotions and creates opportunities seamlessly to tap into listeners’ emotional state through their physiological responses.

Keywords
emotion, mode, music, percussiveness, psychophysiology, tempo
Evaluation case studies

• Gaming
• Sports
• Art
Collaborative immersive game

```c
// Show roman numerals
if (level.sequences[level.currentSequence][level.curDown] == RNUM) {
    if (IsUnicode(level.chalk_hud1))
        level.chalk_hud1 destroy_hud();
    if (IsUnicode(level.chalk_hud2))
        level.chalk_hud2 destroy_hud();
    level.chalk_huds = [];
    index = 0;
    x = 0;
    rounds = level.round_number;
    while (int(rounds / 10) > 0)
```
Sports
Mood Swings: design and evaluation of affective interactive art

LETICIA S. S. BIALOSKORSKI†,†,*, JOYCE H. D. M. WESTERINK†,‡ and EGON L. VAN DEN BROEK†,‡
†User Experience Group, Philips Research Europe, High Tech Campus 34, 5656 AE
Eindhoven, The Netherlands
‡Center for Telematics and Information Technology (CITI), University of Twente, P.O. Box
217, 7500 AE Enschede, The Netherlands

(Received 1 March 2009; final version received 21 June 2009)

The field of affective computing is concerned with developing empathic products, such as affective consumer products, affective games, and affective art. This paper describes Mood Swings, an affective interactive art system, which interprets and visualizes affect expressed by a person. Mood Swings consists of eight luminous orbs that react to movement. When a person experiences certain emotion, his/her movements are claimed to have certain characteristics. Based on the integration of a framework for affective movements and a color model, Mood Swings recognizes affective movement characteristics, and subsequently displays a color that matches the expressed emotion. Mood Swings was evaluated in a museum for contemporary art by 36 museum visitors. The Trajectory of Interaction (ToI) was applied to assess common phases in interacting with Mood Swings, i.e. response,
What can we learn from the studies?

• How to observe users in natural settings?
• Unexpected findings resulting from in the wild studies.
• Develop different data collection and analysis techniques to evaluate UX goals such as challenge and engagement.
• The ability to run experiments on the Internet that are quick and inexpensive using crowdsourcing.
• How to recruit a large number of participants using Mechanical Turk or Google Analytics.
<table>
<thead>
<tr>
<th>Method</th>
<th>Controlled settings</th>
<th>Natural settings</th>
<th>Without users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observing</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Asking users</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Asking experts</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Testing</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modeling</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>The language of evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analytical evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controlled experiment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowdsourcing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecological validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert review or criticism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Field study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formative evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informed consent form</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In the wild evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Living laboratory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictive evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scope</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summative evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usability laboratory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usability testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users or participants</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Participants’ rights and getting their consent

- Participants need to be told why the evaluation is being done, what they will be asked to do and their rights.
- Informed consent forms provide this information.
- The design of the informed consent form, the evaluation process, data analysis and data storage methods are typically approved by a high authority (e.g., medical-ethical committee).
Things to consider when interpreting data

• Reliability: does the method produce the same results on separate occasions?
• Validity: does the method measure what it is intended to measure?
• Biases: Are there biases that distort the results?
• Scope: How generalizable are the results?
Validity

• Content validity
  – Agreement among experts on the domain of interest.
  – The degree to which a feature/parameters of a given signal represents a construct.
  – The degree to which a feature/parameters of a given signal represents all facets of the domain.

• Criteria-related validity: quality of the translation from the preferred measurement to an alternative, rather than to what extent the measurement represents a construct.
  – Predictive: measurements before the event.
  – Postdictive: measurements after the event.
  – Concurrent: reliability metric is applied (in relation to a standard).

• Construct validity: development of a nomological network / ontology / semantic network build around the construct of interest. This requires theoretical grounded, observable, operationable definitions of all constructs and the relations between them.

• Ecological validity refers to the influence of the context on measurements. Notes. Some events are sparse in real life! and Context is often crucial!
Triangulation

“the strategy of using multiple operationalizations of constructs to help separate the construct under consideration from other irrelevancies in the operationalization” (Heath, 2001)

Advantages:
1. Distinct signals can be used to validate each other;
2. Extrapolations can be made based on multiple data sets, providing more certainty.
3. More solid ground, or even a ground truth is obtained for the interpretation of data, as multiple perspectives are used.
Key points

- Evaluation and design are very closely integrated.
- Some of the same data gathering methods are used in evaluation as for establishing requirements and identifying users’ needs; for example, observation, interviews, and questionnaires.
- Evaluations can be done in controlled settings such as laboratories, less controlled field settings, or where users are not present.
- Usability testing and experiments enable the evaluator to have a high level of control over what gets tested, whereas evaluators typically impose little or no control on participants in field studies.
- Things to consider (incl. validity and triangulation)
Chapter 14
Evaluation Studies:
From Controlled to Natural Settings
The aims

• Explain how to do usability testing

• Outline the basics of experimental design

• Describe how to do field studies
Usability testing

• Involves recording performance of typical users doing typical tasks.
• Controlled settings.
• Users are observed and timed.
• Data is recorded on video and key presses and mouse movements are logged.
• Possibly: additional sensors such as eye-tracking and biosignals.
• The data is used to calculate performance times, and to identify and explain errors.
• User satisfaction is evaluated using questionnaires and interviews.
• Field observations may be used to provide contextual understanding.
Experiments & usability testing

- Experiments test hypotheses to discover new knowledge by investigating the relationship between two or more variables.

- *Usability testing is applied experimentation.*

- *Usability testing is an empirical science.*

- Developers check that the system is usable by the intended user population for their tasks.
Usability testing & research

Usability testing

- Improve products
- Few participants
- Results inform design
- Usually not completely replicable
- Conditions controlled as much as possible
- Procedure planned
- Results reported to developers

Experiments for research

- Discover knowledge
- Many participants
- Results validated statistically
- Must be replicable
- Strongly controlled conditions
- Experimental design
- Scientific report to scientific community
Usability testing

• Goals & questions focus on how well users perform tasks with the product.
• Comparison of products or prototypes is common.
• Focus is on time to complete task and the number + type of errors.
• Data collected by video and interaction logging.
• Testing is central.
• User satisfaction questionnaires and interviews provide data about users’ opinions.
Testing conditions

• Usability lab or other controlled space.
• Emphasis on:
  – selecting representative users;
  – developing representative tasks.
• 5-10 users typically selected.
• Tasks usually around 30 minutes
• Test conditions are the same for every participant.
• Informed consent form explains procedures and deals with ethical issues.
Types of data

- Time to complete a task.
- Time to complete a task after a specified time away from the product.
- Number and type of errors per task.
- Number of errors per unit of time.
- Number of times online help and manuals accessed.
- Number of users making an error.
- Number of users successfully completing a task.
How many participants is enough for user testing?

- The number is a practical issue.
- Depends on:
  - schedule for testing;
  - availability of participants;
  - cost of running tests.
- Typically 5-10 participants.
- Some experts argue that testing should continue until no new insights are gained.
Usability lab with observers
Portable labs

Portable Observation Lab

Designed for travelling and ideal for observing people in their natural environment!
Portable equipment
Mobile head-mounted eye tracker

Figure 14.5 The mobile head-mounted eye-tracker
Source: Picture courtesy of SensoMotoric Instruments (SMI), copyright 2010.
Usability testing of the iPad

- 7 participants with 3+ months experience with iPhones
- Signed an informed consent form explaining:
  - what the participant would be asked to do;
  - the length of time needed for the study;
  - the compensation that would be offered for participating;
  - participants’ right to withdraw from the study at any time;
  - a promise that the person’s identity would not be disclosed; and
  - an agreement that the data collected would be confidential and would be available to only the evaluators
- Then they were asked to explore the iPad.
- Next they were asked to perform randomly assigned specified tasks.
Examples of the tasks

<table>
<thead>
<tr>
<th>App or website</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>iBook</td>
<td>Download a free copy of <em>Alice’s Adventures in Wonderland</em> and read through the first few pages.</td>
</tr>
<tr>
<td>Craigslist</td>
<td>Find some free mulch for your garden.</td>
</tr>
<tr>
<td>eBay</td>
<td>You want to buy a new iPad on eBay. Find one that you could buy from a reputable seller.</td>
</tr>
<tr>
<td><em>Time</em> Magazine</td>
<td>Browse through the magazine and find the best pictures of the week.</td>
</tr>
<tr>
<td>Epicurious</td>
<td>You want to make an apple pie for tonight. Find a recipe and see what you need to buy in order to prepare it.</td>
</tr>
<tr>
<td>Kayak</td>
<td>You are planning a trip to Death Valley in May this year. Find a hotel located in the park or close to the park.</td>
</tr>
</tbody>
</table>

*Table 14.1* Examples of some of the tests used in the iPad evaluation (adapted from Budiu and Nielsen, 2010).

Example of the equipment

Figure 14.6 The setup used in the Chicago usability testing sessions
Source: Copyright Nielsen Norman Group, from report available at http://www.nngroup.com/reports/.
Problems and actions

• Problems detected:
  – Accessing the Web was difficult
  – Lack of affordance and feedback
  – Getting lost
  – Knowing where to tap

• Actions by evaluators:
  – Reported to developers
  – Made available to public on nngroup.com

• Accessibility for all users important
Experiments

- Test hypothesis
- Predict the relationship between two or more variables.
  - *Independent variable* is manipulated by the researcher.
  - *Dependent variable* is influenced by the independent variable.
- Typical experimental designs have one or two independent variables.
- Validated statistically and replicable.
Experimental designs

• Different participants - single group of participants is allocated randomly to the experimental conditions.

• Same participants - all participants appear in both conditions.

• Matched participants - participants are matched in pairs, e.g., based on expertise, gender, etc.
## Different, same, matched participant design

<table>
<thead>
<tr>
<th>Design</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different</td>
<td>No order effects</td>
<td>Many subjects &amp; individual differences a problem</td>
</tr>
<tr>
<td>Same</td>
<td>Few individuals, no individual differences</td>
<td>Counter-balancing needed because of ordering effects</td>
</tr>
<tr>
<td>Matched</td>
<td>Same as different participants but individual differences reduced</td>
<td>Cannot be sure of perfect matching on all differences</td>
</tr>
</tbody>
</table>
Content-Based Image Retrieval Benchmarking: Utilizing Color Categories and Color Distributions

Egon L. van den Broek, Peter M. F. Kisters and Louis G. Vuurpilj
Nijmegen Institute for Cognition and Information, Radboud University Nijmegen, THE NETHERLANDS

From a human-centered perspective three ingredients for Content-Based Image Retrieval (CBIR) were developed. First, with their existence confirmed by experimental data, 11 color categories were utilized for CBIR and used as input for a new color space segmentation technique. The complete HSI color space was divided into 11 segments or bins, resulting in a unique CBIR 11 color quantization scheme. Second, a new weighted similarity function was introduced. It exploits within-bin statistics, describing the distribution of color within a bin. Third, a new CBIR benchmark was successfully used to evaluate both new techniques. Based on the 4050 queries judged by the users, the 11 bin color quantization proved to be useful for CBIR purposes. Moreover, the new weighted similarity function significantly improved retrieval performance, according to the users.

Journal of Imaging Science and Technology 49: 293–301 (2005)

Introduction
Digital media are rapidly replacing their analog counterparts. Less than 10 years ago a digital photo camera was solely used in professional environments. In contrast, nowadays many home users own a digital photo camera. This development is accompanied by (i) the increasing number of images present on the internet, (ii) the availability of the internet for an increasing number of people, and (iii) a decline in digital storage costs.

As a result, the need for browsing image collections has emerged. This development gave birth to a new field within Information Retrieval (IR): image retrieval. When images are part of a web page or when images are textually annotated in another form, IR techniques can be utilized. However, how do we search annotated?

We will first discuss quantitative arguments, followed by qualitative arguments that point out the relevance of Content-Based Image Retrieval (CBIR). CBIR uses features of the image itself, i.e., color, texture, shape, and spatial characteristics, which enables us to search for images that are not textually annotated.

Murray determined in his article, “Sizing the internet”, on July 19, 2000, that 2.1 billion unique pages were present on the internet. He further states that “internet growth is accelerating, indicating that the internet has not yet reached its highest growth period.” Currently, estimates of the number of unique pages range from over 50 million up to 8 billion.

Mark on the scale below how good you rate these results as a whole

Figure 3. The interface of a query such as was presented to the subjects. They were asked to select the best matching images and to rate their satisfaction.
Field studies

- Field studies are done in natural settings.
- “In the wild” is a term for prototypes being used freely in natural settings.
- Aim to understand what users do naturally and how technology impacts them.
- Field studies are used in product design to:
  - identify opportunities for new technology;
  - determine design requirements;
  - decide how best to introduce new technology;
  - evaluate technology in use.
Context-aware field data collection

Digital Context Aware Solutions

The Internet of Things
The right information to the right device at the right time
The old days …
about metaphores and other things

Figure 14.8 UbiFit Garden’s glanceable display: (a) at the beginning of the week (small butterflies indicate recent goal attainments; the absence of flowers means no activity this week); (b) a garden with workout variety; (c) the display on a mobile phone (the large butterfly indicates this week’s goal was met).

Data presentation

• The aim is to show how the products are being appropriated and integrated into their surroundings.

• Typical presentation forms include:
  – Vignettes / excerpts (short stories or sketches)
  – Critical incidents,
  – Patterns, and narratives.
Visualization (1)
Visualization (3)

Bad visualizations have …
• too much text,
• excessive ornamentation,
• gaudy colors, and
• clip art.

Good visualizations
• Are not so easy to make …,
• Take into account its users,
• Link to known visualizations, and
• Take into account 10 skills …
Visualization (3)

There are Ten Visualization Skills

- Color
- Size
- Shape
- Number
- Mood
- Perspective
- Background
- Direction
- Movement
- Other Senses

To learn how to teach students visualization skills: Purchase the Beginners and Intermediate Teaching Visualization Power Points

Copyright 2012 © Dr. Erica Warren www.goodsensorylearning.com
Visualization (4)
Key points

• Usability testing takes place in controlled usability labs or temporary labs.
• Usability testing focuses on performance measures, eg. how long and how many errors are made when completing a set of predefined tasks. Indirect observation (video and keystroke logging), user satisfaction questionnaires and interviews are also collected.
• Affordable, remote testing systems are more portable than usability labs. Many also contain mobile eye-tracking and other devices.
• Experiments test a hypothesis by manipulating certain variables while keeping others constant.
• The experimenter controls independent variable(s) in order to measure dependent variable(s).
• Field studies are evaluation studies that are carried out in natural settings to discover how people interact with technology in the real world.
• Field studies that involve the deployment of prototypes or technologies in natural settings may also be referred to as ‘in the wild’.
• Sometimes the findings of a field study are unexpected, especially for in the wild studies in which explore how novel technologies are used by participants in their own homes, places of work, or outside.