Chapter 5 - Evaluation

• Types of Evaluation
  – Formative vs. Summative
  – Quantitative vs. Qualitative
  – Analytic vs. Empirical

• Analytic Methods
  – Cognitive Walkthrough
  – Heuristic Evaluation
  – GOMS and KLM
  – Motor Functions: Fitt's Law, Steering Law

• Empirical Methods
  – Field Studies und Lab Studies
  – Longitudinal und Diary Studies
  – Usability Scales
Formative vs. Summative Evaluation

- Formative: what and how to (re)design
- Summative: how did we do?

• M. Scriven: The methodology of evaluation, 1967
Qualitative vs. Quantitative Evaluation
Analytic vs. Empirical Evaluation

Scriven, 1967: “If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you may just study the kind and speed of the cuts it makes in the hands of a good axeman.”
Empirical and Analytic Methods are Complementary

- Empirical evaluation produces facts which need to be analyzed.
- Analytic evaluation produces facts which need to be tested (empirically).
Chapter 5 - Evaluation

• Types of Evaluation
  – Formative vs. Summative
  – Quantitative vs. Qualitative
  – Analytic vs. Empirical

  • Analytic Methods
    – Cognitive Walkthrough
    – Heuristic Evaluation
    – GOMS and KLM
    – Motor Functions: Fitt's Law, Steering Law

  • Empirical Methods
    – Field Studies und Lab Studies
    – Longitudinal und Diary Studies
    – Usability Scales
Cognitive Walkthrough

...One or more evaluators...

...Step by step...

...along well-defined tasks...

1. Is the correct action for executing the next step always clearly defined? Does the user know what to do next?

2. Is the correct action clearly recognizable? Does the user actually find it?

3. Does the user receive a sufficient feedback after executing the action, such that he can determine whether the action was executed successfully?
10 Usability Heuristics

• Visibility of system status
• Match between system and the real world
• User control and freedom
• Consistency and standards
• Error prevention
• Recognition rather than recall
• Flexibility and efficiency of use
• Aesthetic and minimalist design
• Help users recognize, diagnose, and recover from errors
• Help and documentation
Detailed Checklist Example

http://www.stcsig.org/usability/topics/articles/he-checklist.html

### Visibility of System Status

The system should always keep user informed about what is going on, through appropriate feedback within reasonable time.

<table>
<thead>
<tr>
<th></th>
<th>Review Checklist</th>
<th>Yes No N/A</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Does every display begin with a title or header that describes screen contents?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>Is there a consistent icon design scheme and stylistic treatment across the system?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>Is a single, selected icon clearly visible when surrounded by unselected icons?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>Do menu instructions, prompts, and error messages appear in the same place(s) on each menu?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>In multipage data entry screens, is each page labeled to show its relation to others?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.6</td>
<td>If overtape and insert mode are both available, is there a visible indication of which one the user is in?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.7</td>
<td>If pop-up windows are used to display error messages, do they allow the user to see the field in error?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.8</td>
<td>Is there some form of system feedback for every operator action?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.9</td>
<td>After the user completes an action (or group of actions), does the feedback indicate that the next group of actions can be started?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.10</td>
<td>Is there visual feedback in menus or dialog boxes about which choices are selectable?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
<tr>
<td>1.11</td>
<td>Is there visual feedback in menus or dialog boxes about which choice the cursor is on now?</td>
<td>⬜️ ⬜️ ⬜️</td>
<td></td>
</tr>
</tbody>
</table>
Goals, Operators, Methods & Selection Rules (GOMS)

- **Selection rules**

- **Methods**

- **Operators**

- **Goals**

Card / Moran / Newell: The Psychology of HCI, 1983

Based on Material by A. Butz & A. Krüger
Keystroke Level Model (KLM)

Used times in experimental average:

- **K** (Keystroke): Pressing a key: $t_K = 0.28s$
- **P** (Pointing): Pointing to a position on screen: $t_P = 1.1s$
- **B** (Mouse button): Pressing/releasing mouse button: $t_B = 0.1s$
- **H** (Homing): Switch between keyboard and mouse: $t_H = 0.4s$
- **M** (Mental preparation): Mental preparation of successive operation: $t_M = 1.35s$
- **R(t)** (Response time): Response time of the systems (within $t$ seconds, system-dependent).

Card / Moran / Newell: The Psychology of HCI, 1983
Data according to D. Kieras (http://courses.wccnet.edu/~jwithrow/docs/klm.pdf)
KLM Example

- Which of the methods M1 or M2 is faster?

- **M1**: Switch to mouse, move mouse pointer to file icon, clicking the icon, dragging to trash icon and release, switch to keyboard

- **M2**: Switch to mouse, selecting the icon, switch to keyboard, press ‘delete’

\[
t_{M1} = t_H + t_P + t_B + t_P + t_B + t_H = 0.4 + 1.1 + 0.1 + 1.1 + 0.1 = 2.8s
\]

\[
t_{M2} = t_H + t_P + t_B + t_H + t_K = 0.4 + 1.1 + 0.1 + 0.4 + 0.28 = 2.28s
\]
More Sophisticated KLM table

- **K** - Keystroke (.12 - 1.2 sec; .28 recommended for most users).
  - Expert typist (90 wpm): .12 sec
  - Average skilled typist (55 wpm): .20 sec
  - Average nonsecretarial typist (40 wpm): .28 sec
  - Worst typist (unfamiliar with keyboard): 1.2 sec
- **T(n)** - Type a sequence of n characters on a keyboard (n * K sec).
- **P** - Point with mouse to a target on the display (1.1 sec).
  - The actual time required can be determined from Fitts’ law.
  - For typical situations, it ranges from .8 to 1.5 sec, with an average of 1.1 sec.
- **B** - Press or release mouse button (.1 sec).
- **BB** - Click and release mouse button (.2 sec).
- **H** - Home hands to keyboard or mouse (.4 sec).
Speed vs. Accuracy
Fitts’ Experiment

\[ MT = a + b \times ID = a + b \times \log_2\left(\frac{D}{W} + 1\right) \]
Enlarge Targets, the Right Way!

http://www.particletree.com/features/visualizing-fittss-law/
Not All Pixels Are Equal (before Fitts’ Law)

http://www.particletree.com/features/visualizing-fittss-law/

Infinite Target Widths at Edges

Corners are the easiest places to reach because they have infinite dimensions.
Steering Law ???

Time for Driving Along a Narrow Road

\[ T = a + b \cdot \int_{S} \frac{1}{W(s)} \, ds \]
Narrow Roads on Screens

\[ T = a_1 + b_1 \times \log_2 \left( \frac{nh}{h} + 1 \right) + a_2 + b_2 \times \frac{w}{h} + \ldots \]

vertical: Fitts’ law

horizontal: steering law
Chapter 5 - Evaluation

• Types of Evaluation
  – Formative vs. Summative
  – Quantitative vs. Qualitative
  – Analytic vs. Empirical

• Analytic Methods
  – Cognitive Walkthrough
  – Heuristic Evaluation
  – GOMS and KLM
  – Motor Functions: Fitt's Law, Steering Law

• Empirical Methods
  – Field Studies und Lab Studies
  – Longitudinal und Diary Studies
  – Usability Scales

http://www.amazon.de/dp/0857028294
Quality Properties of Empirical Methods

- Objectivity
- Reproducibility
- Validity
  - internal
  - external
- Relevance
Field Study vs Lab Study

• External Validity
• Internal Validity
• Effort

Source: www.xperienceconsulting.com
Variables and Values

- Nominal
- Ordinal
- Cardinal

Experiment

independent

dependent

http://www.bundesliga.de/
http://www.gebr-clasen.de/-g400/weltkarte_KB.png
http://www.kreativrad.de/img/parts/fahrrad-massanfertigung-korpergroesse.png
Observation Study (Example)

• One independent variable: Participation in tutorials (Yes / No)
  – Assuming participation is voluntary
• One dependent variable: Achieved grade in test
• 108 subjects, 54 “yes”, 54 “no” (to participation question)
• Measurement shows: Grade positively **correlated** with tutorial participation
• Beware of **confounding variables**!
Controlled Experiment

• One independent variable: Participation in tutorials (Yes / No)
  – assigned randomly to subjects !!!
• One dependent variable: Achieved grade in test
• 108 subjects, 54 “participating” condition, 54 “not-participating” condition
• Measurement: Grade positively correlated with participation
• Causal relationship established: Participation in tutorials leads to better grade
Experiment Design

<table>
<thead>
<tr>
<th></th>
<th>HCI1</th>
<th>Analysis</th>
<th>Algebra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Condition 1</td>
<td>Condition 2</td>
<td>Condition 3</td>
</tr>
<tr>
<td>No</td>
<td>Condition 4</td>
<td>Condition 5</td>
<td>Condition 6</td>
</tr>
</tbody>
</table>

- 2 Variables with 2 resp. 3 values: $2 \times 3 = 6$ Conditions
- **within-subjects**: everybody does everything
- **between-groups**: groups, each group does one condition
- Vary the order to avoid **learning and fatigue effects**
  - Randomisation
  - Permutation
  - Latin square

<table>
<thead>
<tr>
<th></th>
<th>Cond. 6</th>
<th>Cond. 1</th>
<th>Cond. 5</th>
<th>Cond. 2</th>
<th>Cond. 4</th>
<th>Cond. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. 5</td>
<td>Cond. 6</td>
<td>Cond. 4</td>
<td>Cond. 1</td>
<td>Cond. 3</td>
<td>Cond. 2</td>
<td></td>
</tr>
<tr>
<td>Cond. 2</td>
<td>Cond. 3</td>
<td>Cond. 1</td>
<td>Cond. 4</td>
<td>Cond. 6</td>
<td>Cond. 5</td>
<td></td>
</tr>
<tr>
<td>Cond. 1</td>
<td>Cond. 2</td>
<td>Cond. 6</td>
<td>Cond. 3</td>
<td>Cond. 5</td>
<td>Cond. 4</td>
<td></td>
</tr>
<tr>
<td>Cond. 4</td>
<td>Cond. 5</td>
<td>Cond. 3</td>
<td>Cond. 6</td>
<td>Cond. 2</td>
<td>Cond. 1</td>
<td></td>
</tr>
<tr>
<td>Cond. 3</td>
<td>Cond. 4</td>
<td>Cond. 2</td>
<td>Cond. 5</td>
<td>Cond. 1</td>
<td>Cond. 6</td>
<td></td>
</tr>
</tbody>
</table>
Hypotheses and Significance

• H: Tutorial participants achieve better grades in test.
• H₀: Tutorial participants and non-participants achieve in average the same grades in test. (*null hypothesis*)
• Effect size = difference of mean values (unknown in advance)
• Trick: Instead of proving H, dis-prove H₀. Then H is implicitly proven – independent of effect size.

• Significance:
  – *p-value*: *probability* of obtaining the observed results when null hypothesis is true
  – *statistical significance*: p-value less than *significance level* Often 0,05 (= 5%)
  – obtaining p-values: *tests* dependent on experiment design
Longitudinal and Diary Studies

![Diary Study Example](http://www.hcii.cmu.edu/M-HCI/2011/BOA-PlanningTools/images/diary_study.jpg)
USE: Usefulness, Satisfaction and Ease of Use

- Lund 2001: 30 questions with 7-point Likert scales

<table>
<thead>
<tr>
<th>USEFULNESS</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It helps me be more effective.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. It helps me be more productive.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. It is useful.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. It gives me more control over the activities in my life.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. It makes the things I want to accomplish easier to get done.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. It saves me time when I use it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. It meets my needs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. It does everything I would expect it to do.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EASE OF LEARNING</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20. I learned to use it quickly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. I easily remember how to use it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. It is easy to learn to use it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23. I quickly became skillful with it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SATISFACTION</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>24. I am satisfied with it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25. I would recommend it to a friend.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26. It is fun to use.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SUS: System Usability Scale

• Brooke (DEC) 1986
  - "Quick and dirty", very popular
  - 10 questions
  - 5-point Likert scale
  - Adapted for Web sites:
    Tullis / Stetson (Fidelity Investments) 2004
NASA TLX

• Measurement for perceived workload
  – NASA AMES Research 1986
  – 100 points per subscale, 5-point steps (i.e. neutral plus 10 values in each direction)

http://humansystems.arc.nasa.gov/groups/TLX/
### PANAS

Positive and Negative Affect Scale

<table>
<thead>
<tr>
<th>Positive Affect</th>
<th>Negative Affect</th>
</tr>
</thead>
<tbody>
<tr>
<td>attentive</td>
<td>upset</td>
</tr>
<tr>
<td>interested</td>
<td>hostile</td>
</tr>
<tr>
<td>alert</td>
<td>irritable</td>
</tr>
<tr>
<td>excited</td>
<td>scared</td>
</tr>
<tr>
<td>enthusiastic</td>
<td>afraid</td>
</tr>
<tr>
<td>inspired</td>
<td>ashamed</td>
</tr>
<tr>
<td>proud</td>
<td>guilty</td>
</tr>
<tr>
<td>determined</td>
<td>nervous</td>
</tr>
<tr>
<td>strong</td>
<td>jittery</td>
</tr>
<tr>
<td>active</td>
<td>distressed</td>
</tr>
</tbody>
</table>

---

*Based on Material by A. Butz & A. Krüger*

User Experience (UX) Design

• Marc Hassenzahl

• “Good UX is the consequence of fulfilling the human needs for autonomy, competency, stimulation (self-oriented), relatedness, and popularity (others-oriented) through interacting with the product or service (i.e. hedonic quality). Pragmatic quality facilitates the potential fulfillment of be-goals.”

• Goal types:
  • Do-goals: Want to send a message through a digital medium
  • Be-goals: Send a message to feel related to another person

• Criteria for usability: change from technical aspects to aspects of human personality

http://hassenzahl.wordpress.com
AttrakDiff

Four dimensions:
• pragmatic quality (PQ)
• hedonic quality - identity (HQ-I)
• hedonic quality - stimulation (HQ-S)
• attractiveness (ATT).

www.attrakdiff.de
AttrakDiff
Visualization

http://attrakdiff.de
Domain-Specific Tests: Automotive Example
Lane Change Task

- Standardized test (ISO 26022)
- Driving situation (primary task)
  - Demands for lane changes at non-predictable times
- Accompanied by secondary task
- Measures attention split primary/secondary task

Based on Material by A. Butz & A. Krüger