Usability IV

Three real-world examples of how systematic usability studies can enhance products
Three Studies

How have other usability researchers and practitioners used metrics in their work?

I. Redesigning a website cheaply and quickly

II. Redesign of the CDC.gov website

III. Drug label design
I. Redesigning a Website Cheaply and Quickly

**Object:** Website of a company in the entertainment news and content industry

**Purpose:** More visitors + motivate them to spend more time

**Background:** More page views → Increased revenues from advertising

**Challenge:**
- Appeal to user and support business needs
- Develop advertising strategies
- Quickly (4 month) and cheaply

**Procedure:** Incrementally, multiphase approach
3 Phases

Phase 1: Testing Competitor Websites

Phase 2: Testing Three Different Design Concepts

Phase 3: Testing a Single Design
Phase 1: Testing Competitor Websites

Learn from the successes and failures of similar websites
→ Conduct a competitive study

First: Discover people’s reaction to concepts
→ Observe the behaviour of 15 participants (Target group)
Sample Tasks

Given: [Movie name]

Task:

- Visit [www.companyname.com]
- Find the most convenient place and time to see [Movie name]
- Find out what TV shows and movies [Actor name] has been in

Purpose:
Identify the good and bad elements of each design
→ Compose the best single design
Questionnaire

Participants should rate the site concerning different characteristics

Please pick a number from the scale to show how well each word or phrase below describes the website.

Strongly Disagree  1  2  3  4  5  6  7  Strongly Agree

__ Credible  __ Fun to use  __ Engaging
__ Easy to use  __ Frustrating  __ Helpful
__ Annoying  __ Boring

Questionnaire used in Phase 1 to measure website appeal.
Outcome: Phase 1

Appeal score of the Website = Total score from the questionnaire

→ Get major usability issues
→ Receive an insight into the (dis-)advantages between other competitive sites
Phase 2: Testing Three Different Design Concepts

• 7 Participants

• Designers created new visual designs

• Including people's comments and reactions and the survey data
The Design Tested (I)

Each participant evaluated 2 of 3 designs:

**Version 1:** Medium graphical treatment + high level of unrelated ads

**Version 2:** Minimal graphical treatment + low level of related and unrelated ads

**Version 3:** High graphical treatment + moderate level of related ads
The Design Tested (II)

Participants tasks:
• Describe their first impression
• Accomplish some tasks on the homepage

Meanwhile:
Team noted user interaction, comments and success score for each task

Finally:
Questionnaire to measure their preferences
# Questionnaire

Participants should rate the site after each evaluation

Please circle the number from the scale to indicate your opinion of the website.

<table>
<thead>
<tr>
<th>Scale</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Dissatisfying to Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Unattractive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Very Credible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Satisfying to Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Attractive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Credible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questionnaire used to measure design preference.
Results: Satisfaction

→ Only a moderate difference
Results: Attractiveness and Credibility

- More attractive design → Higher score in credibility
- User does not mind ads if they are related to the content
- Lower in attractiveness + highest level of unrelated ads → Lowest score in credibility
Negative Reactions

Design 1 had the most negative comments

→ Relevancy of the ads had a stronger effect than the number of ads

→ More negative comments and reactions on unrelated ads
Outcome: Phase 2

- **Interests in trends and design insights**, not in statistical significance
- Only 7 participants, but nevertheless a **strong influence on the team's decision**
- Advertising guidelines (placement, design or types)
Phase 3: Testing a Single Design

- Some **modification** to the respondents preferred design

- Another round of **usability testing** (Fewer participants)

- **Similar tasks**, success scores and user interaction were noted
Results: Phase 3

Satisfaction level rating:

Average satisfaction rating: 5.8 out of 7
Relatively success score: 92%
Outcome: Phase 3

• **Positive results** from this study → Right path

• Tasks can be easily accomplished

• Satisfaction scores a little bit slower as expected (Prototype → e.g., search engine)
Conclusion

• Complete redesign a site in a few short month

• Collected data made team decisions easier

• Only 7 days user testing, only 27 participants → Sufficient data
Questions?
II. Redesign of the CDC.gov Website

- **Object:** Website of the Centers for Disease Control and Prevention (CDC), 2006
- **Purpose:** Redesign + optimize the usability
- **Who:** 6-person User Experience team
Major Usability Activities

- Interviewing users, stakeholders, partners, and web staff
- Conducting a detailed analysis of web and search and call logs
- Conducting parallel design sessions
- ...

PARTICULARLY: **Usability testing**

→ Baseline test
→ First-click tests
→ Final prototype tests
Usability Testing Levels (I)

The testing model includes 5 usability testing levels:

*Level 1:* Traditional inspection evaluations (heuristic evaluations, expert reviews,...)

*Level 2:* Algorithmic reviews with scenarios

BUT:

- **Skip Level 1** due to the weakness of inspection evaluations
- **Do Level 2 at the end** on the final version
Usability Testing Levels (II)

Usability test that are...

**Level 3:** ... moderately controlled +
small number of participants (about 8)

**Level 4:** ... tightly controlled +
only enough participants to make weak inferences to the population

**Level 5:** ... very tightly controlled +
sufficient number of participants to make strong inferences to the population
Overview

Baseline Test

Wireframing and FirstClick Testing

Pilot Test

Final Test

Launch

Monday

Tuesday

Wednesday (last day)
Baseline Test (I)

• Main idea: Collect data you can compare with later on

• Establish the human performance

• Determine the user satisfaction levels

→ Get major usability issues

For the existing website

Baseline Test → Wireframing + FirstClick Test → Pilot Test → Final Test → Launch
Baseline Test (II)

• 68 participants
• 36 scenarios
• Each participant dealt with 10 scenarios

• First: Browse the right answer without using the website's search engine

• Later: “Search-only” test
  → Determining the impact of not allowing users to search

Baseline Test → Wireframing + FirstClick Test → Pilot Test → Final Test → Launch
## Baseline Test (III)

<table>
<thead>
<tr>
<th></th>
<th>Browse</th>
<th>Search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Success rate</strong></td>
<td>54%</td>
<td>49%</td>
</tr>
<tr>
<td><strong>Average time</strong></td>
<td>2.4 min</td>
<td>2.8 min</td>
</tr>
<tr>
<td><strong>Average page views</strong></td>
<td>7.1 pages</td>
<td>7.7 pages</td>
</tr>
<tr>
<td><strong>Satisfaction (out of 100)</strong></td>
<td>46</td>
<td>49</td>
</tr>
</tbody>
</table>

3 minutes per scenario
Otherwise: “Unsuccessful” + starting next scenario

---

**Baseline Test** → Wireframing + FirstClick Test → Pilot Test → Final Test → Launch
Task Scenarios

Change the scenarios with the poorest performance

→ Use the informations gained from the baseline test

→ Focus designers on the homepage issues with the lowest success rate

→ No changes on “good” scenarios
Qualitative Findings (I)

Participants should ...
... type their overall impressions
... specify what they like best and least
... indicate what they would change

Evaluators ...
... note problems participants had
... debriefed them at the end of each test
Qualitative Findings (II)

Observations after the baseline test:

- Too much informations
- A-Z index is a good tool, but hard to find
- Inconsistency in layout, navigation and look and feel
  
→ Invaluable observations!
Wireframing

Baseline Test → **Wireframing** + FirstClick Test → Pilot Test → Final Test → Launch

=================================================================
Wireframing (I)

• Create several competing wireframes

• Combine them in 3 homepage wireframes (A, B and C)

• B was removed after one quick test

→ A and C were tested “head-to-head”

65 participants in 136 scenarios (68 wireframe A, 68 wireframe C) for 1 hour
Baseline Test → Wireframing + FirstClick Test → Pilot Test → Final Test → Launch

HCI SS09 ~ Thomas Dackweiler ~ 4. Juni 2009
FirstClick Testing (I)

FirstClick testing:
Collect and **analyze only the first click** they made after reading the scenario

Background:
- First click mostly very critical
- If the first click is difficult → Frequently problems finding the correct answer

---------------------------------------------------------------------

Baseline Test → Wireframing + **FirstClick Test** → Pilot Test → Final Test → Launch

HCI SS09 ~ Thomas Dackweiler ~ 4.Juni 2009
FirstClick Testing (II)

Results:
No reliable difference between A and C in terms of success

But:
C had the faster performance

Poor success in baseline test → Poor performance in the FirstClick test
FirstClick Testing (III)

Baseline Test → Wireframing + **FirstClick Test** → Pilot Test → Final Test → Launch

---

HCI SS09  ~  Thomas Dackweiler  ~  4.Juni 2009
**Prelaunch Test**

**Combine** the best design decisions on both wireframes **to a final wireframe**
→ Basis for developing a graphic prototype

Prelaunch usability test (2 parts):
- Pilot test
- Final test

Baseline Test → Wireframing + FirstClick Test → **Pilot Test** → Final Test → Launch

HCI SS09  ~  Thomas Dackweiler  ~  4.Juni 2009
Pilot Test

18 participants and 56 scenarios each with 3 categories:
- 26 FirstClick from the Homepage
- 24 FirstClick from one of the second-level pages
- 9 “homepage to content page”

→ **Summarizing and analyzing** the data

→ **Changes** were made **to the homepage** and a few **to the test**
Final Test

Final test:
19 participants and 56 revised scenarios

Baseline Test → Wireframing + FirstClick Test → Pilot Test → **Final Test** → Launch
================================================================================================
"Homepage to Content Page"

Navigation through the website to find information (→ Baseline test)
→ Observe improvement from the baseline test to the final prelaunch test

<table>
<thead>
<tr>
<th>Topic</th>
<th>Baseline Test</th>
<th>Prelaunch Test</th>
<th>%-Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success rate</td>
<td>62</td>
<td>78</td>
<td>26%</td>
</tr>
<tr>
<td>Average time</td>
<td>96</td>
<td>81</td>
<td>19%</td>
</tr>
<tr>
<td>Average page views</td>
<td>8,3</td>
<td>4,9</td>
<td>69%</td>
</tr>
<tr>
<td>Satisfaction (out of 100)</td>
<td>46</td>
<td>78</td>
<td>70%</td>
</tr>
</tbody>
</table>
Conclusion (I)

Three main uses of usability testing metrics:

1) Establish an **overall baseline success rate** for comparison

2) **Identify** scenarios where users have the **most problems**
   → Look for patterns

3) **Identify** scenarios with consistently **high levels of human performance**
Conclusion (II)

Procedure:
Changes were made after each test → Improved performance? → Test again → ...

All in all:
• The **usability testing program was very valuable** in the CDC.gov revision of the homepage

• High-level usability testing is absolute necessary
Questions?
III. Measuring the Effects of Drug Label Design and Similarity on Pharmacists' Performance

- **Object**: Standard template for Abbott's drug labels

- **Purpose**: Reduce review time of new drugs
  - + anticounterfeiting feature to the label
  - + more cost effective

- **Who**: User Centric (Research firm)

- **Procedure**: Investigate applications and elements of the new label
Initial studies

New label → Better search efficiency and information-processing efficiency

→ Locate and identify critical drug information more quickly
New Label

New Abbott drug label means:

• Change in label design
• Increase in consistency across labels
Main objective

Has the new label an **impact** on **drug cartons rather than drug bottles**?

Yes? → Why?

• Different label design?
• Increased similarity across all Abbott labels?
• Combination of these both?
About the study

• Conducted in Europe (Cartons much more common than bottles)

• 20 participants (14 w, 6 m) with pharmaceutical experience

• Typical drug selection tasks

• Ideal metric: **Error rate**
About the study

• Also measured:
  Factors that can **contribute to errors** (Difficult in findings,...)

  → Measuring of ...
    ... the **number of eye fixation** prior to target selection (Search efficiency)

    ... **average fixation duration** (Information processing difficulty)

    ... **pupil diameter** (Cognitive workload)

    ... **response times**
Stimuli

Drugs were stored in drawers:

- 16 cartons arranged in a **4 x 4 grid**
- With **8 unique drugs**
- **Multiple dosage** strengths (3 types)
Stimuli

- 4 designs tested
- Each with high+low similarity
Procedure (I)

Given: A picture of a drawer

Task: Locate a drug (e.g., “Find Medotil 500 mg)

Action:
- Read the instructions
- Press the space bar, drawer appears
- Use D, F, K and L on the keyboard to indicate the column
Procedure (II)

- 8 trials per “low” drawer, 8 trials per “high” drawer
- All Abbott labels as target
  - 4 Abbott labels as target, 4 non-Abbott labels
- Order of the trials was randomized
- Drug positions stayed constant
Analysis

Data collected:

• Error rate
• Time-on-task
• 3 eye-related measures:
  Pupil diameter, fixation count, average fixation duration
Fixation Count

Each Abbott label in the low-similarity:

More (avg.) fixations than in the high-similarity

→ Much higher target-to-Abbott label ratio (4 targets/ 5 Abbott labels, 8/16)
→ Verifying look at the instructions
Results and Discussions (I)

- 9 seconds to complete the task (on avg.)

- 5.9% errors ( Mostly wrong key )

- Pupil size was independent → Workload consistent

- New template required fewer fixations (0.88 fixations per trial instead of 1.1) → Better search efficiency
  Closer proximity of the label elements → Smaller scan area

- New template required longer fixations (M = 279 ms instead of 220 ms) → Smaller font and higher information density
Results and Discussions (II)

Overall:

• New labels did **not worse** than the existing labels

• But: **Not show much improvement**
  
  → Possible reason: No experience with those (fictitious) drugs

→ Study with a lot of **scientific aspects**, but only a **few convincing results**
Results and Discussions (II)

Recommendations:

• Longer sessions or multiple sessions over a period of time → Learning and expert performance

• Extend to physical drawers and cartons
Conclusion

What did we extract of these 3 studies?

• You don't need much participant to get sufficient and qualitative data

• Testing – Modification Iteration to achieve improvements

• Baseline tests for comparison

• Deal with the problems weigh most heavily, abstract away from the rest
Thank you
for your attention!