Closing the Loop: From Analysis to Design

Scott Klemmer
TAs: Marcello Bastea-Forte, Joel Brandt, Neil Patel, Leslie Wu, Mike Cammarano

15 November 2007
Analyzing the results

- Quantitative data, which might include:
  - success rates
  - time to complete tasks
  - pages visited
  - error rates
  - ratings on a satisfaction questionnaire

- Qualitative data, which might include:
  - notes of your observations about the pathways participants took
  - notes about problems participants had (critical incidents)
  - notes of what participants said as they worked
  - participants' answers to open-ended questions

Source: Usability.gov
Using the Test Results

• Summarize the data
  • make a list of all critical incidents
    • positive & negative
  • include references back to original data
  • try to judge why each difficulty occurred

• What does data tell you?
  • UI work the way you thought it would?
    • users take approaches you expected?
  • something missing?
Using the Results (cont.)

- Update task analysis & rethink design
  - rate severity & ease of fixing CIs
  - fix both severe problems & make the easy fixes
- Will thinking aloud give the right answers?
  - not always
  - if you ask a question, people will always give an answer, even if it has nothing to do with facts
  - try to avoid specific questions
Measuring Bottom-Line Usability

- Situations in which numbers are useful
  - time requirements for task completion
  - successful task completion
  - compare two designs on speed or # of errors

- Ease of measurement
  - time is easy to record
  - error or successful completion is harder
    - define in advance what these mean

- Do not combine with thinking-aloud. Why?
  - talking can affect speed & accuracy
Analyzing the Numbers

- Example: trying to get task time $\leq 30$ min.
  - test gives: 20, 15, 40, 90, 10, 5
  - mean (average) = 30
  - median (middle) = 17.5
  - looks good!

Factors contributing to our uncertainty:
- small number of test users ($n = 6$)
- results are very variable (standard deviation = 32)

std. dev. measures dispersal from the mean
Analyzing the Numbers (cont.)

- This is what statistics is for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
## Web Usability Test Results

<table>
<thead>
<tr>
<th>Participant #</th>
<th>Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number of participants</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>30.0</td>
</tr>
<tr>
<td>median</td>
<td>17.5</td>
</tr>
<tr>
<td>std dev</td>
<td>31.8</td>
</tr>
</tbody>
</table>

\[
\text{standard error of the mean} = \frac{\text{std dev}}{\sqrt{\text{#samples}}}
\]

\[
\text{typical values will be mean } \pm 2 \times \text{standard error} \quad \Rightarrow 4 \text{ to } 56!
\]

\[
\text{what is plausible? } = \text{confidence (alpha=5%)}
\]

\[
\Rightarrow 95\% \text{ confident between } 5 \text{ & } 56
\]
Analyzing the Numbers (cont.)

- This is what statistics is for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
- Usability test data is quite variable
  - need lots to get good estimates of typical values
  - 4 times as many tests will only narrow range by 2x
    - breadth of range depends on sqrt of # of test users
  - this is when online methods become useful
    - easy to test w/ large numbers of users
Measuring User Preference

- How much users like or dislike the system
  - can ask them to rate on a scale of 1 to 10
  - or have them choose among statements
    - “best UI I’ve ever…”, “better than average”…
  - hard to be sure what data will mean
    - novelty of UI, feelings, not realistic setting …
- If many give you low ratings -> trouble
- Can get some useful data by asking
  - what they liked, disliked, where they had trouble, best part, worst part, etc. (redundant questions are OK)
Reporting the Results

- Report what you did & what happened
- Images & graphs help people get it!
- Video clips can be quite convincing
CASE STUDY
David Akers
evaluation of
Google SketchUp
Study Goals

1. What individual differences (previous software used, computer use, spatial reasoning ability, etc.) best predict performance on simple modeling tasks?

2. What usage log metrics (e.g. frequency of undo operations, frequency of camera operations, etc.) best predict performance on simple modeling tasks?

3. What specific problem do novice SketchUp users encounter most frequently on simple modeling tasks?
n = 54

90% students
35% architecture
20% computer science
10% mechanical engineering
10% civil engineering
25% other (art, physics, etc.)

41% never used
44% novice
15% intermediate
Study Design

1. Entry questionnaire (5 min.)
2. Mental rotation test (15 min.)
3. Video tutorials (15 min.)
4. Free exploration (10 min.)
5. Tasks (3 x 15 min.)
6. Exit questionnaire (5 min.)
Data Size

Event log data (450 MB)
Screen capture video (75 GB!)
3D models (100 MB)
Questionnaires (17 KB)
Study Goals

1. What individual differences (previous software used, computer use, spatial reasoning ability, etc.) best predict performance on simple modeling tasks?

2. What usage log metrics (e.g. frequency of undo operations, frequency of camera operations, etc.) best predict performance on simple modeling tasks?

3. What specific problem do novice SketchUp users encounter most frequently on simple modeling tasks?
Log Analysis of Tool Usage

<table>
<thead>
<tr>
<th>UseTool</th>
<th>Arc</th>
<th>CameraOrbit</th>
<th>CameraZoom</th>
<th>CoordinateSystem..</th>
<th>Erase</th>
<th>MakeComponent</th>
<th>MatchPhoto</th>
<th>Measure</th>
<th>Move/Copy</th>
<th>Paste</th>
<th>Pencil</th>
<th>PositionTexture</th>
<th>Push/Pull</th>
</tr>
</thead>
</table>

Time (minutes)

```
0  50 100 150 200 250 300 350 400 450
```
## Undo Rates

<table>
<thead>
<tr>
<th>Tool</th>
<th>Undo Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Push/Pull</td>
<td>0.54</td>
</tr>
<tr>
<td>Arc</td>
<td>0.55</td>
</tr>
<tr>
<td>Dimension</td>
<td>0.34</td>
</tr>
<tr>
<td>FollowMe</td>
<td>0.28</td>
</tr>
<tr>
<td>Measure</td>
<td>0.26</td>
</tr>
<tr>
<td>Rectangle</td>
<td>0.20</td>
</tr>
<tr>
<td>Erase</td>
<td>0.18</td>
</tr>
<tr>
<td>Circle</td>
<td>0.16</td>
</tr>
<tr>
<td>Pant</td>
<td>0.14</td>
</tr>
<tr>
<td>Pencil</td>
<td>0.12</td>
</tr>
<tr>
<td>Paste</td>
<td>0.10</td>
</tr>
<tr>
<td>Polygon</td>
<td>0.08</td>
</tr>
<tr>
<td>Offset</td>
<td>0.06</td>
</tr>
<tr>
<td>RotateObject</td>
<td>0.04</td>
</tr>
<tr>
<td>Move/Copy</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*UndoRate per Tool Use*
Developer Hypotheses (Wrong)

For the Push/Pull tool:

90% of undo operations are caused by bugs in SketchUp.

10% of undo operations are caused by difficulties with inferencing.
In Analysis mode, statechart and recorded video are synchronized and each can be used to access the other. *Inset:* simultaneous interaction with statechart and video editing is possible on a dual-screen workstation.

**Figure 3**

**d.Tools physical prototyping**
**captures user tests**

Challenges (1/8 – from Grudin)

- Disparity of Work and Benefit
  Groupware applications often require additional work from individuals who do not perceive a direct benefit from the use of the application
Challenges (2/8)

- Critical Mass and Prisoner’s Dilemma
  Groupware may not enlist the “critical mass” of users required to be useful, or can fail because it is never to any one individual’s advantage to use it.
Group Calendaring
Challenges (3/8)

- Disruption of Social Processes
  Groupware can lead to activity that violates social taboos, threatens existing political structures, or otherwise demotivates users crucial to its success
Challenges (4/8)

- **Exception Handling**
  Groupware may not accommodate the wide range of exception handling and improvisation that characterizes much group activity
Challenges (5/8)

- **Unobtrusive Accessibility**
  Features that support group processes are used relatively infrequently, requiring unobtrusive accessibility and integration with more heavily used features.
Challenges (6/8)

- **Difficulty of Evaluation**
  The almost insurmountable obstacles to meaningful, generalizable analysis and evaluation of groupware prevent us from learning from experience.
Track Changes
Challenges (7/8)

- Failure of Intuition
  Intuitions in product development environments are especially poor for multiuser applications, resulting in bad management decisions and error-prone design process.
Challenges (8/8)

- The adoption process
  Groupware requires more careful implementation in the workplace than product developers have confronted.
The Communicator
Eye to the future: iRoom