Introduction

- Usability issues generally thought of as purely qualitative
  - Description of problem users experienced
  - Assessment of underlying cause
  - Recommendations for remedying problems

- Most professionals don’t strongly associate metrics with usability issues
  - Gray area in identifying issue
  - Identifying issue part of iterative process, adds no value

- Chapter reviews simple metrics for usability issues
What Is a Usability Issue?

- Examples of Usability Issues
  - Anything that prevents task completion
  - Anything that takes someone off course
  - Anything that creates some level of confusion
  - Anything that produces an error
  - Not seeing something that should be noticed
  - Assuming something should be correct when it is not
  - Assuming a task is complete when it is not
  - Performing the wrong action
  - Misinterpreting some piece of content
  - Not understanding the navigation

- Usability Finding (positive)
  - Supporting the user in completing a complex transaction without any confusion and in the most efficient way possible
  - Anticipating a user's needs at every step of a process
  - Education a user without any effort involved
  - Displaying complex information in a clear, simple format that users can easily understand

Goal: issues that point to possible improvement in product

“The mental model of the application does not match the user’s mental model”

Usability Finding (positive)

“Many participants were confused by the top-level navigation menu, often jumping around from one section to another, trying to find what they were looking for.”

What Is a Usability Issue?

- Real Issues versus False Issues
  - Was the participant’s behavior thought process, perception, or decision during the task logical?
  - Yes, then it may be a usability issue even if 1 user encountered it
  - No, then likely to be idiosyncratic and can ignore

User clicked on a link that started down the wrong path. Ask, why did you click on the website link?

Wording of the link made it seem like a reasonable place to start

Because it was in front of me

Issue that most, if not all, participants encounter

Participant selected wrong option from poorly worded menu, taken down wrong path, spent significant amount of time looking for target in wrong part of the application

1 out of 10 participants express some confusion around a specific piece of content or terminology on a website

1 out of 12 participants don’t notice something they should have noticed
How to Identify an Issue

- Many ways to identify issues
  - During a study in which you are directly interacting with participant
  - Over the phone using remote testing technology
  - Automated techniques such as online study

- When are issues identified?
  - Predicted beforehand and tracked during a test
  - Surprise issues may appear during testing

How to Identify an Issue

- In-Person Studies
  - Think-aloud protocol verbalize their thoughts while working through tasks
  - May observe the following which point to usability issues
    - Verbal expression of confusion, frustration, dissatisfaction, pleasure, or surprise
    - Verbal expression of confidence or indecision about a particular action that may be right or wrong
    - Participants not saying or doing something they should have done or said
    - Nonverbal behaviors such as facial expressions and/or eye movements

- Automated Studies
  - Requires careful data collection
  - Allow participants to enter verbatim comments at a page or task level
  - Allow participants to enter verbatim comments when a low score is provided (e.g. ease of use)
    - More pointed questions lead to more specific actionable comments
How to Identify an Issue

- When Issues Begin
  - Initial confusion or deviation away from ideal behavior

- When Issues End
  - With task failure
  - When another issue begins
    - Different problem arises and takes place of original issue
    - Sufficiently different issue
  - Recovery from issue
    - User identifies source of confusion
  - No longer relevant

Severity Ratings

- Not all issues are the same
  - Annoying or frustrating
  - Make wrong decisions
  - Loose data

- Severity Rating
  - Help focus attention on issues that really matter

- Many classifications, most simplify to two types
  - Severity rating systems based purely on the user experience
  - Bring in multiple dimensions or factors
Severity Ratings

- Severity Rating Based on the User Experience
  - Easy to implement
  - Provides useful information
  - Usually have 3-5 levels

<table>
<thead>
<tr>
<th>Severity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Any issue that annoys or frustrates participants but does not play a role in task failure.</td>
</tr>
<tr>
<td>Medium</td>
<td>Any issue that contributes to but does not directly prevent task failure.</td>
</tr>
<tr>
<td>High</td>
<td>Any issue that directly leads to task failure.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Minimal error</td>
</tr>
<tr>
<td>4</td>
<td>Minor but irritating</td>
</tr>
<tr>
<td>3</td>
<td>Moderate: waste of time but no loss of data</td>
</tr>
<tr>
<td>2</td>
<td>Severe problem causing loss of data</td>
</tr>
<tr>
<td>1</td>
<td>Catastrophic error, causing loss of data or damage to hardware/software</td>
</tr>
</tbody>
</table>

Severity Ratings

- Severity Rating Based on a Combination of Factors
  - Nielsen based on impact on the user experience coupled with frequency of use and/or impact on the business goals

- Rubin combines into critically score
  - Assigns severity rating on 4-point scale
  - Assigns frequency of occurrence rating on a 4-point scale
  - Adds scores to get critically score between 2 and 8

**Critically Score = Severity Rating + Frequency of Occurrence**

<table>
<thead>
<tr>
<th>Severity Rating</th>
<th>Frequency of Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 = irritant</td>
<td>1 = occurs &lt;10 percent of the time</td>
</tr>
<tr>
<td>2 = moderate</td>
<td>2 = occurs 11 to 50 percent of the time</td>
</tr>
<tr>
<td>3 = severe</td>
<td>3 = occurs 51 to 89 percent of the time</td>
</tr>
<tr>
<td>4 = unusable</td>
<td>4 = occurs more than 90 percent of the time</td>
</tr>
</tbody>
</table>
Severity Ratings

- Using a Severity Ratings System
  - Be consistent
  - Clearly communicate what each level means
  - Try to have more than one usability specialist assign severity ratings to each issue

- Not everyone believes in severity ratings
  - Let audience provide own ratings, they don’t truly understand relative importance
  - Little agreement between usability specialists
  - Severity rating increases with issues personally identified with (evaluator effect)

- Severity rating not perfect, still provide useful purpose

Analyzing and Reporting Metrics for Usability Issues

- Straightforward to identified and prioritize usability issues

- What about analysis of the issues themselves?
  - Depends on the type of usability questions

- Possible usability questions
  - How is the overall usability of the product?
  - Is the usability improving with each design iteration?
  - Where should I focus my efforts to improve the design?

- Analysis performed can be done with or without severity ratings
Analyzing and Reporting Metrics for Usability Issues

- Frequency of Unique Issues
  - Count the unique issues (not frequency of issues)
    - Provides insight into how usability is changing with each iteration
    - May not be significantly better, what if issues remaining are much bigger?
  - Include severity ratings
    - Most telling data is the change in the number of high priority issues

Analyzing and Reporting Metrics for Usability Issues

- Frequency of Issues per Participant
  - Look at number of issues each participant encountered
  - Expect to see number decrease over each design iteration
Analyzing and Reporting Metrics for Usability Issues

- Frequency of Participants
  - Percentage of participant that encountered a specific issue
  - Useful when you focus on whether a specific design element is improved
  - Important that criteria for identifying issues across users and designs are consistent

Issues by Category
- Where to focus design improvements?
- Useful to aggregate issues into broad categories

Issues by Task
- Identify which tasks lead to the most issues
Analyzing and Reporting Metrics for Usability Issues

- Important to reporting positive issues
- Designers and developers should be able to consider you an ally
  - Present only problems, hard to build goodwill
- Document positive issues so they can be propagated to other areas of the product
- Enhances your credibility
  - Need to appear neutral

Consistency in Identifying Usability Issues

- Little agreement on what a usability issue is or how important issue is
  - Group of people who worked together previously separated into two teams
  - Both teams evaluated same website independently with real users
- Result
  - 38% of the usability issues were reported by both teams
  - 32% of the usability issues only reported by Team A
  - 30% of the usability issues only reported by Team B
- What about severity?
  - Looking only at high-severity issues, 88% of these issues shared
Bias in Identifying Usability Issues

- Participants
  - Different level of expertise, domain knowledge, and motivation
  - Different levels of comfort in lab setting

- Tasks
  - Clearly defined, open-ended, self-generated
  - Defined tasks indicate which parts of prototype are exercised and how they are exercised

- Method
  - Traditional lab testing or some type of expert review
  - Length of sessions, thinking aloud, how and when you probe

- Artifact
  - Interaction varies based on paper prototype, functional or semi-functional prototype, or production system

- Environment
  - Direct interaction with user, indirect interaction via conference call or behind one-way mirror

- Moderators
  - Usability professional have different experience, domain knowledge, and motivation

---

Bias in Identifying Usability Issues

- Study on the source of biases
  - Analyzed reports from 9 teams
  - Looked at number of participants in each test (5 to 15)
  - Looked at number of tasks used (5 to 14)
  - Looked at number of usability issues reported (20 to 50)

- Total of 106 unique problems categorized as serious or critical
  - No significant correlation between number of users and percentage of usability problems found
  - Found significant correlation between number of tasks used and percentage of new problems found
Number of Participants

- Hotly debated issues – two camps have emerged
  - Five participants is enough to identify most of the usability issues
  - Nowhere near five participants is enough to identify most of the usability issues

- Five Participants is Enough
  - Majority or 80% issues observed with the first five users
  - Average probability that with each participant about 31% of the usability problems are being observed
    - Assume all issue have equal probability of detection

1 user – 30% problems detected
3 users – 66% problems detected
5 users – 83% problems detected

Seminal paper by Nielsen and Landauer (1993) defined the “magic number 5”
Based on anecdotal evidence from 11 different studies

---

Number of Participants

- Five Participants is Not Enough
  - Magic number 5 challenged by some researchers

- Spool and Schroeder (2001)
  - Participants asked to buy various products at 3 different websites
  - Only 36% of the usability issues discovered after first five users
  - Mainly discounted because individual differences among participants not accounted for

- Lindgaard and Chattratchart (2007)
  - Compared results from two teams
  - Team A uncovered 42% and Team H uncovered 43% percent
  - Team A had 6 users, Team H had 12 users
  - Found only 28% in common, more than 70% of the problems uncovered by only one of the teams
Number of Participants

- Our (Book) Recommendation
  - Five participants per significantly different class of user is usually enough

- Experience seems to support magic number five, but following conditions
  - Scope of evaluation is limited
    - 5 to 10 tasks and about 20 to 30 web pages
  - User audience is well defined and represented
    - If audience known, well represented in testing
    - If more than one unique audience, strive to have 5 participants from each group
    - Challenge is in knowing when you really do have different user groups

Number of Participants

- Calculating p, or probability of detection
  - Line up all issues discovered during test
  - For each user mark how many were observed with that user
  - Add total number of issues identified with each user, divide by total number of issues
  - Overall probability rate for test - take average for all test participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Issue 1</th>
<th>Issue 2</th>
<th>Issue 3</th>
<th>Issue 4</th>
<th>Issue 5</th>
<th>Issue 6</th>
<th>Issue 7</th>
<th>Issue 8</th>
<th>Issue 9</th>
<th>Issue 10</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PB</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PC</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PD</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PE</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PF</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PG</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PH</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PI</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>PJ</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Average proportion $p = 0.49$

How many users needed to identify a certain percentage of issues?

$1 - (1 - p)^n$

where $n$ is the number of users

What proportion of issues would be identified by a sample of three users?

$= 1 - (1 - 0.49)^3$
$= 1 - (0.51)^3$
$= 1 - 0.133$
$= 0.867$ or 87% of issues
Summary

- Easiest way to identify usability issues is during an in-person lab study
  - Can be done in automated using verbatim comments

- Real vs. False Issues
  - Consistent story behind user’s thought process and behavior

- Severity of issues
  - Impact on user experience, frequency of use, impact on business, persistence
  - Based on simple high/medium/low rating
  - Numerous other scales also available

- Common ways to measure usability issues
  - Frequency of unique issues
  - Percentage of participants who experience a specific issue
  - Frequency of issues for different tasks or categories of issues

- Identifying usability issues
  - Consistency
  - Acknowledge where biases may arise