Squeezing 1,000 Users into the Lab, Or, How to Conduct Online User Experience Studies

Bill Albert, Design and Usability Center, Bentley University

Donna Tedesco, Fidelity Investments

Tom Tullis, Fidelity Investments
Agenda

- Overview (Bill)
- Designing a study (Donna)
- Analyzing the data and choosing a tool (Tom)
- Q&A (All)
A caveat (and a plug)

- We can only give you a taste of online usability testing
- Everything you need is in our new book!
Another caveat

- We are big fans of traditional usability testing!
What is online usability testing?
What do you call this?

Unmoderated Usability
Online Usability
Remote Usability
Large-Scale Usability
Self-Guided Usability
Automated Usability
Asynchronous Usability
Missing Piece in our Toolkit

Research methods by **Data Source** vs. **Approach** vs. **Context of Product Use**

- **Data Source**
  - Behavioral
    - Usability Lab Studies
    - Ethnographic Field Studies
  - Attitudinal
    - Participatory Design
    - Focus Groups
    - Phone Interviews

- **Approach**
  - Qualitative (direct)
  - Mix
  - Quantitative (indirect)

- **Context of Product Use** during data collection
  - Natural use of product
  - De-contextualized / not using product
  - Scripted (often lab-based) use of product
  - Combination / hybrid

© 2008 Christian Rohrer
Big Picture
This is becoming more important

- Questions from senior management are becoming more complex
- Pure usability is no longer enough
- Push to measure the UX
- Convergence with market research and web analytics to paint a more complete picture of the UX
5 basic ingredients

1. Interactive system (usually a website)
2. A lot of participants (n>50 per segment)
3. An online tool to moderate study and collect data
4. Tasks (not just attitudes/opinions)
5. Capturing data about their experience
Types of Studies

- Comprehensive evaluation
- UX benchmark
- Competitive evaluation
- Live site vs. prototype comparison
- Feature/function test
- Discovery (or true intent)
Why should you care?

- What are the usability issues, and how big?
- Which design is better, and by how much?
- How do customer segments differ?
- What is the overall satisfaction with using the product?
- How does past experience correlate with ease of use?
- Where and why does abandonment occur?
- What is the user experience like?
Strengths

1. Comparing products
2. Measuring user experience
3. Finding the right participants
4. Focusing on design improvements
5. No moderator biases
Limitations

1. Not well suited to rapid, iterative design
2. Need a deep understanding of issues
3. Studies that require long sessions
4. Lose control over prototypes
5. Internet access
Myths

1. Only test with websites
2. It is very expensive
3. Only gather quantitative data
4. A lot of noise in the data
5. Does not work well as part of the design phase

http://johnnyholland.org/2010/04/09/debunking-the-myths-of-online-usability-testing/
## Complements to lab testing

<table>
<thead>
<tr>
<th>Lab first, then Online</th>
<th>Online first, then Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify/fix “low hanging fruit”, then focus on remaining tasks with large sample size</td>
<td>Identify the most significant issues online through metrics, then use lab study to gather deeper qualitative understanding of those issues</td>
</tr>
<tr>
<td>Generate new concepts, ideas, questions through lab testing, then test/validate online</td>
<td>Collect video clips or more quotes of users to help bring metrics to life</td>
</tr>
<tr>
<td>Validate attitudes/preferences observed in lab testing</td>
<td>Gather all the metrics to validate design - if it tests well, then no need to bring users into the lab</td>
</tr>
</tbody>
</table>
Planning a Study

- Choosing a tool
- Recruit strategy
- Sampling strategy
- Incentives
- Timeline

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Plan Goals and Structure of Study</td>
<td>01/04/2010</td>
<td>01/06/2010</td>
<td>3d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Build prototypes (2 Wizard Flows, with slight difference)</td>
<td>01/05/2010</td>
<td>01/11/2010</td>
<td>5d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Build tasks and questions into existing online study tool</td>
<td>01/07/2010</td>
<td>01/11/2010</td>
<td>3d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Pilot and Test tool (2 iterations)</td>
<td>01/11/2010</td>
<td>01/13/2010</td>
<td>3d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Launch Study</td>
<td>01/14/2010</td>
<td>01/15/2010</td>
<td>2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Analysis</td>
<td>01/18/2010</td>
<td>01/19/2010</td>
<td>2d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Present Results</td>
<td>01/20/2010</td>
<td>01/20/2010</td>
<td>1d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Design - Introducing the Study

- Purpose of the Study
- Sponsor/Contact Info (if applicable)
- Time estimate
- Incentive
- Technical requirements
- Legal information/consent
- Instructions
Screening/Starter Questions

- Screening vs. sorting (quotas)
  - Misrepresentation Checks (E.g., brand of pet food)

- Product/Computer/Web Experience
  - Clear and focused; having a product doesn’t imply using it
  - Frequency of use (experience) doesn’t always mean expertise

- Don’t “show your cards” – Distractor questions
Task Construction

- Can have open-ended studies – true intent/intercept, open web

- These tips involve targeted tasks that require providing an answer
  - Will not go over the basics of task construction; just specific to online studies
Task Construction

- Make the task easy to understand
  - No SAT questions
  - End state is clear
  - But not easy to guess

“Find out in what time period the Apollo program was started”

“1960’s”
“1961”
“Kennedy’s Administration”

“In what year was the Apollo program started?”
Task Construction

Some options for collecting task completion:

1) Discrete answer – ISBN number, product price, year, etc.
2) If using a prototype, can tag pages – code or letter, or link to a dummy page with a code or letter
3) Word task to get participant to a page/section and offer page/section titles as answer choices
4) Rely on clickstream data to validate success (preset in tool or after the fact) or ask for URL/title open-ended
5) Self-reported task completion
Task: You want to know how much the total value of your brokerage account has changed since the last close. Were you able to find this information?

- Yes, I found the information
- I'm not sure whether I found it
- No, I could not find the information
Task Construction

Multiple-choice answers

- Easier to analyze than free-form text (answer formats, misspelling, mistyping)
- Consider whether participants can see the answers simultaneously or after-the-fact

Task: You are interested in the Health Sector. Find all 5-Star (as rated by Morningstar) mutual funds in the Health sector. What fund matching these criteria has the highest 1 year performance?

- HealthCare UltraSelect Profund (HCPIX)
- BlackRock Health Sciences (SHSHX)
- Fidelity Select Medical Equipment and Systems Portfolio (FSMEV)
- Live Oak Health Sciences (LOGSX)
- Fidelity Select Pharmaceuticals (FPVHK)
- None of the Above
Task Construction

“None of the Above” Option
- Discourages answers by deduction
- Don’t need to include too many answer choices, because NoA option is possible

“Don’t Know/Give Up” Option
- Can encourage people not to put in too much effort
- ...But:
  - May lend to more natural behavior, keeps people from exiting study altogether if frustrated, good opportunity for follow-up question, “What prompted you to move on from this task?”
Task Construction

Configure the evaluation into logical groupings by separating them into blocks of sections.
Sections can be set up to show a random subset and randomized sequencing, on both the sections in blocks and the blocks themselves.
Click the line between sections to split the sections into separate blocks or to merge them together.

Sections

**Block A**
- 1 Introduction
- 2 Background Questions
- 3 Log into account
- 4 Task 2 - MFE
- 5 Task 3 - Stock Chart
- 6 Task 4 - Stock Recommendation
- 7 Task 5 - Cost Basis_Gain/Loss
- 8 Task 6 - IRA Max contribution
- 9 Task 7 - Expense Ratio of Mutual Fund
- 10 Task 8 - Email Address
- 11 Task 12 - Balances self-report

**Block B**
- 12 Prospect Task Intro
- 13 Task 10 - ToA
- 14 Task 11 - Commissions
- 15 Post-Study

Blocks

Show: All Blocks
Sequence: In Order
Post-Task Questions

Self-reported subjective ratings

- “I found this task to be very easy” or “This task was very easy ---- very difficult”

Open-ended responses

- Post-Task: “How was this task?”
- Conditional: “What did you find difficult about this task?”
Post-Session/Study Questions

- Overall Assessment Tools
  - SUS, CSUQ, QUIS, WAMMI, NPS
    - We like SUS. Short and simple, free, yields reliable results even with smaller numbers of participants

- Open-ended
  - Were there aspects you found challenging?
  - Were there aspects you found intuitive?
  - What would you do to improve this website?

- Broader user-experience questions (the touchy feely-s)
  - “I was inspired by this website”
Demographic Questions & Wrap-Up

- Demographic questions
  - Used to segment data/results
  - Like to place at the end of the survey if possible
    - Less apprehensive, more invested
  - Keep basic, avoid sensitive questions if possible
    - Careful how you ask it – e.g. rather than exact age, provide age ranges.

- Thank You page – warm fuzzies and other housekeeping
Data Analysis

- You can’t “watch over the shoulder” of your participants.
- The analysis will depend upon the goals of the study, but you commonly have:
  - Performance Data
  - Self-reported Data
  - Clickstream Data
  - Verbatim comments
Sample Study: Apollo Program
Performance Data

Percent Correct by Task

Overall Percent Correct
(Error bars represent 95% confidence interval)

Mean Time per Task (mins)
(Error bars represent 95% confidence interval)
Performance Data (cont.)

Breakdown of Task Completion Status

- NASA
  - 58% Correct
  - 24% Incorrect
  - 18% Gave Up

- Wikipedia
  - 71% Correct
  - 17% Incorrect
  - 11% Gave Up

Efficiency (% Correct / Time in mins)

- NASA
  - Efficiency: 30% (±10%)

- Wikipedia
  - Efficiency: 80% (±10%)
Self-Reported Data

Task Ease Rating (1=Very Difficult, 5=Very Easy)

Overall Rating of Ease of Finding Info

Overall Rating of Visual Appeal
Self-Reported Data (cont.)

Mean SUS Scores

Frequency Distribution of SUS Scores

% of Participants

SUS Score
Clickstream Data
Clickstream Data (cont.)

Percent of Participant Tasks That Used Search

# Pages Beyond Minimum

Incorrect Tasks
Correct Tasks
Choosing An Online Study Tool

- Four commercial tools that we will discuss:
  - Loop11
  - RelevantView
  - UserZoom
  - WebEffective

- Our criteria for selecting the tools were that they had to support:
  - Testing of live websites or functional prototypes
  - Presentation of tasks to the participants
  - Automated collection of data, including at least some task metrics (e.g., task success, time, ratings).
  - Collection of clickstream data.

- Or, you can give up some functionality and “roll your own” online study.
WebEffective 7.0 has been released.
WebEffective 7.0 features a highly interactive clickstream diagramming tool that researchers can use to analyze participant browsing behavior, e.g., to identify the most common paths followed by successful participants. All active studies have been resumed.
Criteria for Choosing a Tool

- **Cost**
  - What is the cost per study?
  - Is there a cost per participant?
  - Are there any extra charges for certain data (e.g., clickstream data)?
  - Do you offer a subscription model (e.g., for unlimited studies or participants within a time period)?

- **Participants and Recruiting**
  - How do you normally handle recruiting of participants?
  - Do you have your own panel of potential participants? If so, what are the charges for using it? How large is it?
  - Will you work with other panel providers?
  - How does recruiting work if we provide our own participants?

- **The Product Being Tested**
  - Are there any restrictions on what can be tested?
  - Can you test live sites as well as prototypes?
  - Does the test require any code or other changes to the site?
  - Can the site be inside a firewall (e.g., an intranet site)?
  - Can the test just deliver tasks and questions and have the user testing something else (e.g., a phone-based system)?

- **The Online Study Tool—for the Researcher**
  - Is a website provided for setting up the study and accessing the data?
  - Are there limits on the number of studies you can have at one time?
  - Can different users be set up to use the online tool? Can they have different roles or permissions?
  - Is it possible to save a study as a template for re-use? Or is it possible to copy a previous study’s design?
  - Is it possible to easily preview or pilot the study?
  - Once the study is live, is it possible to make any changes?

- **The Online Study Tool—for the Participant**
  - Does the tool require the participants to download anything to their computer? Can they still participate without downloading it?
  - If anything is downloaded, what is done with it after the study is over (e.g., is it removed)?
  - Does the tool provide a progress indicator?
  - What browsers and versions does the tool support? Has it been tested on all of them?
  - Does the tool require a minimum screen resolution?
  - Does the tool work on mobile devices (e.g., iPhone)?
  - Can participants access the study instructions again any time during the study (e.g., via a link in the taskbar)?
Criteria (cont.)

■ Study Setup
  ■ Do you provide assistance in setting up a study? Are there extra charges?
  ■ Can the order of answer options for a multiple-choice question be randomized?
  ■ Can logic be used to skip questions depending upon previous answers?
  ■ Are there limits on the number of tasks or questions?
  ■ Is there flexibility in the placement of the taskbar?
  ■ Can “welcome” and “thank you” messages be fully customized?
  ■ Can participants be redirected to a custom URL after the study?
  ■ Can parameters (e.g., a participant ID) be passed to the study, or from the study to a follow-up URL (e.g., for an additional survey)?
  ■ Can rules be set up to fill participant quotas based on screening criteria (e.g., if you want an equal mix of men and women, to stop accepting one if the quota is met)?

■ Tasks
  ■ Can task order be randomized? Can one or more tasks be fixed (e.g., at the beginning or end)?
  ■ Are user-generated tasks supported (where the participant defines the task)?
  ■ Can the participant see the task description and the site at the same time?
  ■ How does the participant indicate task completion?
  ■ Is an option for giving up on the task provided?
  ■ Can the task answer either be multiple-choice or open-ended? If multiple choice, is an “Other” option, with an input field, provided?
  ■ Can tasks be chosen based on answers provided by the participant (e.g., about areas of the site they do or don’t use)?
  ■ Can tasks be randomly selected for each participant from a larger set?

■ Task Data Collected
  ■ How is task success determined (e.g., by reaching one or more predefined URLs, by the answer to a question)?
  ■ Is task time collected? If so, what all does it include (e.g., time to read the task, answer the task question, etc)?
  ■ Can the participant provide comments on any task?
  ■ Are rating scales or other questions supported for each task? How many and what questions can be asked before or after each task?
  ■ Is clickstream data collected? Is it tied to the specific tasks?
Criteria (cont.)

- **Self-reported Data Collected**
  - What kinds of rating scales are supported (e.g., Likert scales, semantic differential)?
  - Are standard sets of questions supported (e.g., the System Usability Scale)?
  - Can the number of scale values on a rating scale be controlled?
  - What types of questions are supported (e.g., single-choice radio buttons, multiple-choice check boxes, drop-down lists, open response fields, etc)?

- **Data Analysis**
  - Is an online tool provided for analysis of the data?
  - Is it possible to access the data while the study is still “live” or does the study have to be closed?
  - Can all of the data be exported or downloaded? What formats are supported (e.g., Excel, Word)?
  - Is the data time-stamped (e.g., in case the site changes mid-study and you need to segment the data appropriately)?
  - What visualizations are provided for task success data, task times, and self-reported data (e.g., bar graphs)?
  - Are cross-tabs supported (e.g., looking at the task success data split by younger vs. older participants)?
  - If clickstream data is supported, what visualizations and other analysis methods are provided for it (e.g., click heatmaps, analysis of most successful pages)?
  - Are mechanisms provided for easily filtering out data for participants who weren’t really trying (e.g., “speeders” and “cheaters”)?
Another Option: Rolling Your Own

- A way of building an online usability study based on online survey tools.
- Sample study built using the free version of SurveyGizmo. Allows you to collect:
  - For each task:
    - Task success
    - Task time
    - Subjective rating
    - Comments
  - Overall comments or ratings, such as the System Usability Scale (SUS).
  - Any demographic data needed.
- But you don’t get any clickstream data.
Sample Study of MeasuringUX.com

Measuring the User Experience

A companion website to the book by Tom Tullis and Bill Albert.

Including articles, spreadsheets, links, and other resources related to usability metrics. We're adding to this site regularly, so please check back!

Articles & Presentations

- Results of Online Usability Study Comparing Obama and McCain Websites (Nov. 14, 2008)
- "Word Clouds" from Measuring the User Experience (Nov. 8, 2008)
- Tips and Tricks for Measuring the User Experience (PDF). Presentation by Tom Tullis and Bill Albert at the UPA Boston Usability and User Experience 2008 Conference (May 28, 2008)
- A Timestamp Macro for Task Timing (March 29, 2008)
- Calculating a Confidence Interval for Task Completion using the Adjusted Wald Method (March 28, 2008)
- Results of Online Usability Study of Apollo Program

News and Announcements

- We will be teaching the following upcoming tutorials: NEW
  - 90-minute Course at the CHI 2009 Conference:
    Tips and Techniques for Measuring the User Experience
    Boston, MA, USA
    April 8, 2009
  - Full-day Tutorial at the UPA 2009 Conference:
    Usability Metrics 101
    Portland, Oregon, USA
    June 9, 2009
- Bill Albert has been named as the new Director of the Design and Usability Center at Bentley University. NEW
- Check out the profile of Bill in the February 2009 issue of the UPA Voice. NEW
- Forthcoming book: Online Usability Testing: Improving the User Experience through Automated Studies

Tools and Services

http://www.webusabilitystudy.com/MeasuringUX.htm
Rolling Your Own

For details see:
http://www.measuringux.com/RollingYourOwnOnlineStudy.ppt
http://www.measuringux.com/SurveyGizmo.htm
Thank You!

**Contact Information:**

- Bill Albert: walbert@bentley.edu
- Donna Tedesco: donna.tedesco@fmr.com
- Tom Tullis: tom.tullis@fmr.com