

What's Better? Moderated Lab Testing or Unmoderated Remote Testing?

Susan Fowler

You're probably very familiar with moderated in-lab usability testing, and you may have tried a few unmoderated tests—an online questionnaire, for example. However, full-scale unmoderated tests may be less familiar. For overviews of remote unmoderated testing, in-lab testing, and moderated remote testing, come to this talk. You will hear about the advantages and disadvantages of the two approaches in terms of statistical validity, data quality and quantity, expense, timeframes, finding qualified subjects, and more.

About the Speaker:

Susan Fowler has been an analyst for Keynote Systems, Inc., which offers remote unmoderated user-experience testing. She is currently a consultant at FAST Consulting and an editorial board member of User Experience magazine. With Victor Stanwick, she is an author of the Web Application Design Handbook (Morgan Kaufmann Publishers).

Toe to Toe: What's Better? Moderated Lab Testing or Unmoderated Remote Testing?

Susan Fowler
FAST Consulting
718 720-1169
susan@fast-consulting.com

What's in this talk

- Definitions & differences between moderated and unmoderated tests
- What goes into a remote unmoderated test script?
- What goes into the remote-study report?
- Comparisons between moderated and unmoderated tests
- References

Definition of Terms

- **Moderated:** In-lab studies and studies using online conferencing software with a moderator. Synchronous.
- **Unmoderated:** Web-based studies using online tools and no moderator. Asynchronous.

Differences

- **Rewards:**
 - *Moderated*—\$50 cash, gifts.
 - *Unmoderated*—\$10 online gift certificates, coupons or credits, raffles.
- **Finding participants:**
 - *Moderated*—use a marketing/recruiting company or a corporate mail or email list.
 - *Unmoderated*—send invitations to a corporate email list, intercept people online, or use a pre-qualified panel.

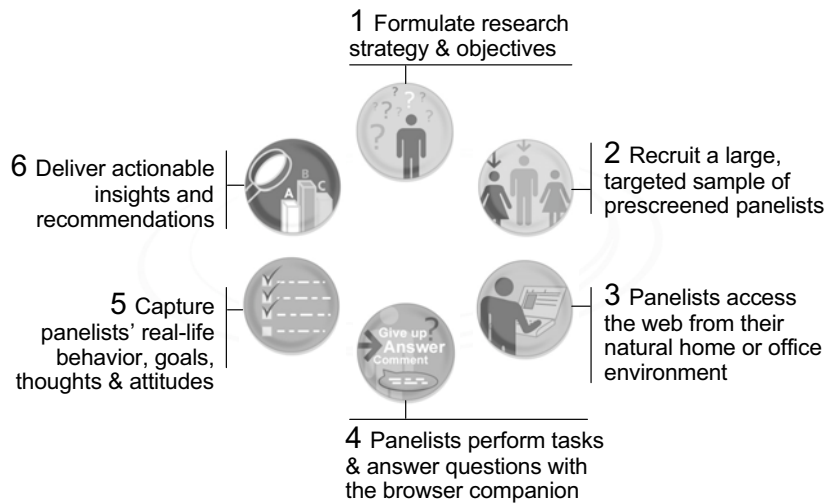
Differences

- **Qualifying participants:**
 - *Moderated*—ask them; have them fill in a questionnaire at start.
 - *Unmoderated*—ask them in a screener section and knock out anyone who doesn't fit (age, geography, disease, etc.).

Differences

- **Test scripts:**
 - *Moderated*—the moderator has tasks he or she wants the participant to do, and the moderator and the notetakers track the questions and difficulties themselves.
 - *Unmoderated*—the script contains both the tasks and the questions that the moderator wants to address.

How Keynote Systems' Tool Works



Creating an Unmoderated Test Script

- **Screener**
- **For each task: "Were you able to...?"**
 - Ask "scorecard" questions--satisfaction, ease of use, organized
 - Ask "what did you like?" and "what did you not like?"
 - Provide a list of frustrations with an open-ended "other" option at end.
- **Wrap-up:**
 - Overall scorecard, "would you return," "would you recommend," email address for gift

Analyzing Unmoderated Results

- **Quantitative data:** Satisfaction, ease of use, and organization scorecards, plus other Likert results, are analyzed for statistical significance and correlations
- **Qualitative data:** Lots and lots of comments (responses to open-ended questions)
- **Clickstream data:** Where did the participants actually go? First clicks, standard paths, fall-off points

Comparisons

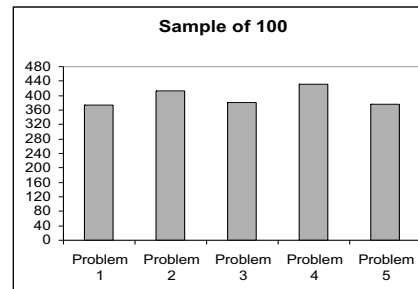
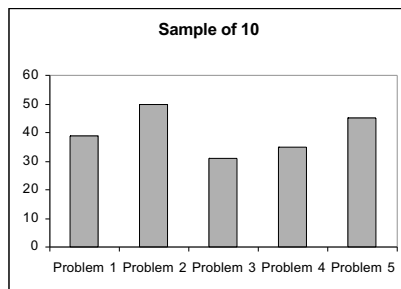
- Statistical validity
- Shock value of participants' comments
- Quality of the data
- Quantity of the data
- Missing information
- Cost
- Time
- Subjects
- Environment
- Geography

Comparisons: Statistical validity

- What's the real difference between samples of 10 (moderated) and 100 (unmoderated)?
 - "The smaller number is good to pick up the main issues, but you need the larger sample to really validate whether the smaller sample is representative.
 - "I've noticed the numbers swinging around as we picked up more participants, at the level between 50 and 100 participants. At 100 or 200 participants, the data were completely different." *Ania Rodriguez, ex-IBM, now Keynote director*

Comparisons: Statistical validity

- It's just math...



Comparisons: Statistical validity

- What's the real difference between samples of 10 (moderated) and 100 (unmoderated)?
 - "In general, quantitative shows you where issues are happening. For *why*, you need qualitative."
 - But "to convince the executive staff, you need quantitative data."
 - "We also needed the quantitative scale to see how people were interacting with eBay Express. It was a new interaction paradigm [faceted search]—we needed click-through information, how deep did people go, how many facets did people use? *Michael Morgan, eBay usability group manager; uses UserZoom & Keynote*

Comparisons: Shock value

- Are typed comments as useful as audio or video in proving that there's a problem?
- *Ania*:
 - "Observing during the session is better than audio or video. While the test is happening, the CEOs can ask questions. They're more engaged."
 - That being said, "You can create a powerful stop-action video using Camtasia and the clickstreams."

Comparisons: Shock value

- Are typed comments as useful as audio or video in proving that there's a problem?
- *Michael:*
 - "The typed comments are very useful—top of mind. However, they're not as engaging as video." So, in his reports, he combines qualitative Morae clips with the quantitative UserZoom data.
 - "We also had click mapping (heat maps and first clicks)," and that was very useful. "On the first task, looking for laptops, we found that people were going to two different places."

Comparisons: Shock value

- My experience regarding responses:
 - A few said that the ads looked too much like a registration or job application (a few panelists complained about having to register, never realizing that the registration screen was part of an ad).
 - "When I did the search it didn't bring up job search results, but an promotion to register for a product."
 - "I didn't like how it took me to a "special offer" first, before showing me the results."
 - "Absolutely HATE the ad screen for the lending company between the front page and the results window. At first I thought the site itself was asking me to register before realizing it was a solicitation. It came across as at best misleading, at worst deceptive."
 - "A sign-up for something else happened after I hit the search key and it confused me whether that was part of what I had to do to search or not."
 - "You can't just look for a job, you have to be registered."

Comparisons: Quality of the data

- Online and in the lab, what are the temptations to be less than honest?
 - In the lab, some participants want to please the moderator.
 - Online, some participants want to steal your money.

Comparisons: Quality of the data

- How do you prompt participants to explain why they're stuck if you can't see them getting stuck?
 - The questions in the script help. In the task debriefing, include a general set of explanations from which people can choose. For example, "The site was slow," "Too few search results," "Page too cluttered."
 - Let people stop doing a task, but ask them why they quit.

Comparisons: Quantity of data

- What is too much data? What are the trade-offs between depth and breadth?
 - “I’ve never found that there was too much data. I might not put everything in the report, but I can drill in 2 or 3 months later if the client or CEO asks for more information about something.”
 - With more data, “I can also do better segments” (for example, check a subset like “all women 50 and older vs. all men 50 and older). —*Ania Rodriguez*

Comparisons: Quantity of data

- What is too much data? What are the trade-offs between depth and breadth?
 - “You have to figure out upfront how much you want to know. Make sure you get all the data you need for your stakeholders.
 - “You won’t necessarily present all the data to all the audiences. Not all audiences get the same presentation.” The nitty-gritty goes into an appendix.
 - “You also don’t want to exhaust the users by asking for too much information.” —*Michael Morgan*

Comparisons: Missing data

- What do you lose if you can't watch someone interacting with the site?
 - Some of the language they use to describe what they see. "eBay talk is 'Sell your item' and 'Buy it now.' People don't talk that way. They say, 'purchase an item immediately.'" –*Michael Morgan*
 - Reality check. "The only way to get good data is to test with 6 live users first. We find the main issues and frustrations, and then we validate them by running the test with 100 to 200 people." –*Ania Rodriguez*
 - Body language, tone of voice, and differences because of demographics

Comparisons: Relative expense

- What are the relative costs of moderated vs. unmoderated tests?
 - What's your experience?

Comparisons: Time

- Which type of test takes longer to set up and analyze: moderated or unmoderated?
 - What's your experience?

Comparisons: Subjects

- Is it easier or harder to get qualified subjects for unmoderated testing?
 - Keynote and UserZoom offer pre-qualified panels.
 - If you want to pick up people who use your site, an invitation on the site is perfect.
 - If you do permission marketing and have an email list of customers or prospects already, you can use that.
- How do you know if the subjects are actually qualified?
 - Ask them to answer screening questions. Hope they don't lie. Don't let them retry (by setting a cookie).

Comparisons: Environment

- In unmoderated testing, participants use their own computers in their own environments. However, firewalls and job rules may make it difficult to get business users as subjects.
- Also, is taking people out of their home or office environments ever helpful—for example, by eliminating interruptions and distractions?

Comparisons: Geography

- Remote unmoderated testing makes it relatively easy to test in many different locations and time zones.
- However, moderated testing in different locations may help the design team understand the local milieu better.

References & Links

- Keynote Systems demo: "Try it now" on http://keynote.com/products/customer_experience/web_ux_research_tools/webeffective.html
- UserZoom: <http://www.userzoom.com/index.asp>
- Farnsworth, Carol. (Feb. 2007) "Using Quantitative/Qualitative Customer Research to Improve Web Site Effectiveness." http://www.nycupa.org/pastevent_07_0123.html
- Tullis, T. S., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, M. (2002) An Empirical Comparison of Lab and Remote Usability Testing of Web Sites (PDF). Usability Professionals Association Conference, July 2002, Orlando, FL. (<http://members.aol.com/TomTullis/prof.htm>)

Recommendations: Statistics

- Darrell Huff, "How to Lie With Statistics," W. W. Norton & Company (September 1993) http://www.amazon.com/How-Lie-Statistics-Darrell-Huff/dp/0393310728/ref=pd_bbs_sr_1/102-0663507-0637745?ie=UTF8&s=books&qid=1190492483&sr=1-1
- Julian L. Simon, "Resampling: The New Statistics," 2nd ed., October 1997, <http://www.resample.com/content/text/index.shtml>
- Michael Starbird, "What Are the Chances? Probability Made Clear & Meaning from Data," The Teaching Company, <http://www.teach12.com/store/course.asp?id=1475&p=c=Science%20and%20Mathematics>

Contact us anytime!

- Susan Fowler has been an analyst for Keynote Systems, Inc., which offers remote unmoderated user-experience testing. She is currently a consultant at FAST Consulting and an editorial board member of *User Experience* magazine. With Victor Stanwick, she is an author of the *Web Application Design Handbook* (Morgan Kaufmann Publishers).
- 718 720-1169; cell 917 734-3746
- <http://fast-consulting.com>
- susan@fast-consulting.com