L26: Web-Scale Usability Testing

- GR5 out, due May 11 (by demonstration to your TA)
In 1988, the USS Vincennes guided missile cruiser shot down an Iranian airliner over the Persian Gulf with almost 300 people aboard. The cruiser was already embroiled in a battle with Iranian gunboats, so it was already in a battle footing, but two UI failures also contributed to this incident. The radar operator interpreted the airliner as an F-14, descending as if to attack, rather than (in reality) a civilian plane that was climbing after takeoff. Both failures seemed to be caused by user interface. The IFF system was reporting the signal from an Iranian F14 on the ground at an airfield hundreds of miles away, not the signal from the airliner; and the plane’s altitude readout showed only its current altitude, not the direction of change in altitude, leaving to the operator the mental comparison and calculation to determine whether the altitude was going up or down. (Peter Neumann, “Aegis, Vincennes, and the Iranian Airbus“, Risks v8 n74, May 1989).
Today’s lecture is about usability evaluation in the brave new world of the Web. The Web enables experiments on a larger scale, for less time and money, than ever before. Web sites with millions of visitors (such as Google, Amazon, Microsoft) are capable of answering questions about the design, usability, and overall value of new features simply by deploying them and watching what happens. The trick lies in how to conduct those experiments. Today’s lecture will discuss some of the latest practices in online experimentation.

Let’s start with an example. Here are two versions of a web page, for a site that sells customized reports about sex offenders living in your area. The goal of the page is to get visitors to fill out the yellow form and buy the report. Both versions of the web page have the same information; they just present it in different ways. In fact, the version on the right is a revised design, which was intended to improve the design by using two fat columns, so that more content could be brought “above the fold” and the user wouldn’t have to do as much scrolling.

We could look closely at these examples and pick them apart with respect to usability principles (visibility, learnability, efficiency, etc.), and the designers were doubtlessly thinking about principles and justifications for the design decisions they made. But at the end of the day, which design is more effective for the end goal of the web site – converting visitors into sales?

The designers answered this question by conducting an experiment. Half the users to their web site were randomly assigned to see one version of the page, and the other half saw the other version. The users were then tracked to see how many of each actually filled out the form to buy the report. In this case, the revised design actually failed – 244 users bought the report from the original version, but only 114 users bought the report from the revised version.

The important point here is not which aspects of the design caused the failure (which I frankly don’t know, because a variety of things changed in the redesign). The point is that the web site conducted a randomized experiment and collected data that actually tested the revision. That’s not the same as just rolling out the revised version and seeing what happens – there’s a subtle but important difference. This kind of experiment is often called an A/B test.

Source: http://www.alistapart.com/articles/designcancripple
Here’s another example – a shopping cart for a web site. Again, a number of changes have been made between the left side (the original version) and the right side (the revised version). When this redesign was tested with an A/B test, it produced a startling difference in revenue – users who saw the cart on the left spent ten times as much as users who saw the cart on the right! The designers of this site explored further and discovered that the problem was the “Coupon Code” box on the right, which led users to wonder whether they were paying too much if they didn’t have a coupon, and abandon the cart. Without the coupon code box, the revised version actually earned more revenue than the original version.
One more example. At the end of every page in Microsoft’s online help (e.g. for Word and Excel) is the question on the left, asking for feedback about the article. If you press any of the buttons, it displays a textbox asking for more details.

A proposed revision to this interface is shown on the right. It was motivated by two arguments: (1) it gives more fine-grained quantitative feedback than the yes/no question; and (2) it’s more efficient for the user, because it takes only one click rather than the minimum two clicks of the left interface.

When these two interfaces were A/B tested on Microsoft’s site, however, it turned out that the 5-star interface produced an order of magnitude fewer ratings – and most of them were either 1 star or 5 stars, so they weren’t even fine-grained.
The term “A/B testing” actually comes from marketing. Other fields have other names for the idea – in the context of usability studies in the lab, we’ve been calling them controlled experiments. The setup is basically the same: you choose an independent variable (like the UI design) with at least two alternatives to test; you choose a dependent variable that you’re going to use to measure the difference between those alternatives.

The distinction in web-based A/B testing is that your web site **automatically and randomly** assigns users to a condition.
Ramp-up

• A/B testing can be risky
  – you’re doing your testing with real users on a deployed system
  – so bugs have real consequences
• Don’t go to 50/50 ratio between Control and Treatment immediately
  – Ramp up slowly: first 99.9% / 0.1%, then 99%/1%, etc.
Assigning Users to Conditions

- Use hashing to partition users
  - MD5 hash of (user id, experiment name) => 128-bit value
  - split the 128-bit space into Control and Treatment
  - for rampup, initially the partition is unbalanced (e.g. 99% / 1%); gradually shift the split point until you reach 50/50

- Why is this better than random number generation?
  - Doesn’t require storing the random assignment
  - Can be done independently by different servers
Power Analysis

• How many users do I need for significance?
  – If the experiment involves too few users, then it may fail to reject the null hypothesis even though it’s false
  – **Power**: probability of correctly rejecting the null hypothesis when it’s false
  – Number of users you need depends on:
    • power desired (typically 80-90%)
    • number of conditions
    • variance of the dependent variable
    • effect size: how much of a difference in dependent variable you care about for decision making
    • statistical test you’re using

• Number of users required determines running time
  – Based on the visit rate of your web site
A/A Tests

- An “experiment” that divides users into two groups with the **same** condition for both groups
  - Good for testing the experimentation infrastructure
  - You shouldn’t see any difference between the groups
    - But wait! If you run 20 A/A tests and test them at the 5% significance level, then on average one of the tests will show a (phantom) significant difference
  - A/A tests also allow estimating the variance of the dependent variable
    - which is useful for power calculations
Ethics: A/B testing never asks the user’s permission to be involved in the test, and doesn’t get informed consent. What do you think about that?

Predictability: when a user visits the web site, things might (randomly) be different. What’s the effect of that?

Numbers, but no explanations: as we saw in our examples at the beginning of the lecture, you get data about how a new design affected bottom-line indicators, but you don’t really find out why. One solution to that is to break down a design with several changes into a few experiments, testing changes individually. Another is to complement large-scale A/B testing with small-scale user testing in the lab, where you have the advantage of think-aloud protocols.

Short-term vs. long-term: a typical A/B test runs only for days or weeks, while the real effect of a new design might be seen only over a long term, as users learn how to use it well. But it’s worth noting that even days or weeks is a longer term than a typical lab-based user study, which might last at most a few hours.
Remote Usability Testing

- Remote synchronous testing
  - using webcam, audio, remote desktop connection
  - shown to be just as effective as face-to-face
- Remote asynchronous testing
  - Approach 1: user identifies and reports critical incidents themselves
    - like bug reporting, but for usability problems
    - users slow down by 3x and report only half as many problems as trained observers would
  - Approach 2: install instrumentation in the web site to track a user’s actions
    - e.g. userfly.com
    - shows details of interaction, but lacks think-aloud and insight into user’s goals and intentions

Recruiting Users Online

- Craigslist is a good source for lab subjects
  - for MIT subjects, the freemoney mailing list also helps
- Mechanical Turk is a labor market for tiny online tasks
  - e.g. paying $0.01 to give some keywords for an image
  - increasingly used by HCI researchers and social scientists to recruit users
- Google AdWords is another way to get users
  - generates high flow, and possibly high cost

Summary

- A/B testing offers fast, accurate testing of new web site designs in actual deployment
- Remote usability testing is getting there
- Web makes it much easier to recruit users than ever before