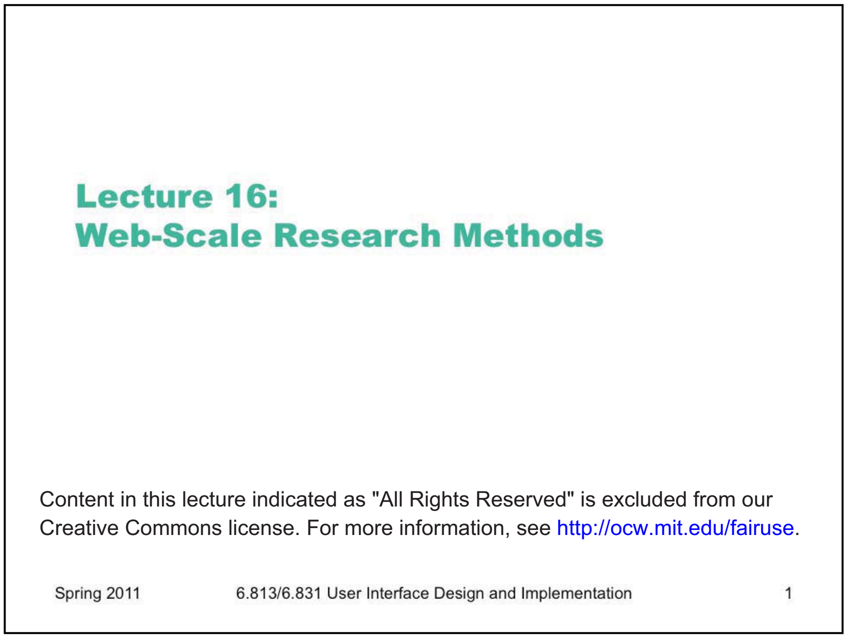
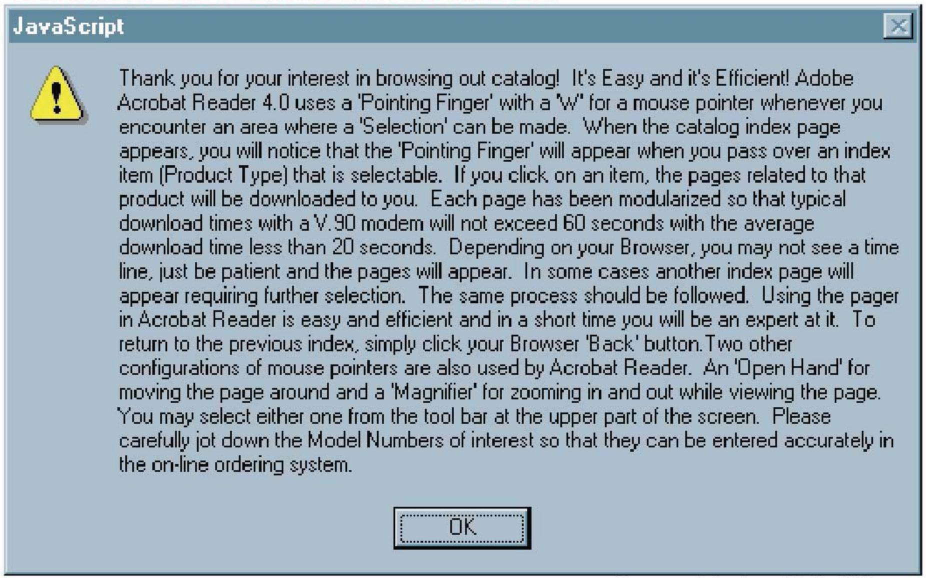


Lecture 16:
Web-Scale Research Methods

UI Hall of Fame or Shame?



Thank you for your interest in browsing out catalog! It's Easy and it's Efficient! Adobe Acrobat Reader 4.0 uses a 'Pointing Finger' with a 'W' for a mouse pointer whenever you encounter an area where a 'Selection' can be made. When the catalog index page appears, you will notice that the 'Pointing Finger' will appear when you pass over an index item (Product Type) that is selectable. If you click on an item, the pages related to that product will be downloaded to you. Each page has been modularized so that typical download times with a V.90 modem will not exceed 60 seconds with the average download time less than 20 seconds. Depending on your Browser, you may not see a time line, just be patient and the pages will appear. In some cases another index page will appear requiring further selection. The same process should be followed. Using the pager in Acrobat Reader is easy and efficient and in a short time you will be an expert at it. To return to the previous index, simply click your Browser 'Back' button.Two other configurations of mouse pointers are also used by Acrobat Reader. An 'Open Hand' for moving the page around and a 'Magnifier' for zooming in and out while viewing the page. You may select either one from the tool bar at the upper part of the screen. Please carefully jot down the Model Numbers of interest so that they can be entered accurately in the on-line ordering system.

Source: Interface Hall of Shame

Once upon a time, this bizarre help message was popped up by a website (Midwest Microwave) when users requested to view the site’s product catalog. The message appears before the catalog is displayed. Clearly this message is a patch for usability problems in the catalog itself. But the message itself has a lot of usability problems of its own! How many problems can you find? Here are a few:

•Overwhelming the user with detail. What’s important here, and what isn’t? (**minimalist design**)

•Horrible layout: no paragraphs, no headings, no whitespace to guide the eye (**aethestic design**)

•No attempt to organize the material into chunks so that it can be scanned, to find out what the user doesn’t already know (**visibility**)

•This information is useless and out of context before the user has seen the task they’ll be faced with (**help and documentation**)

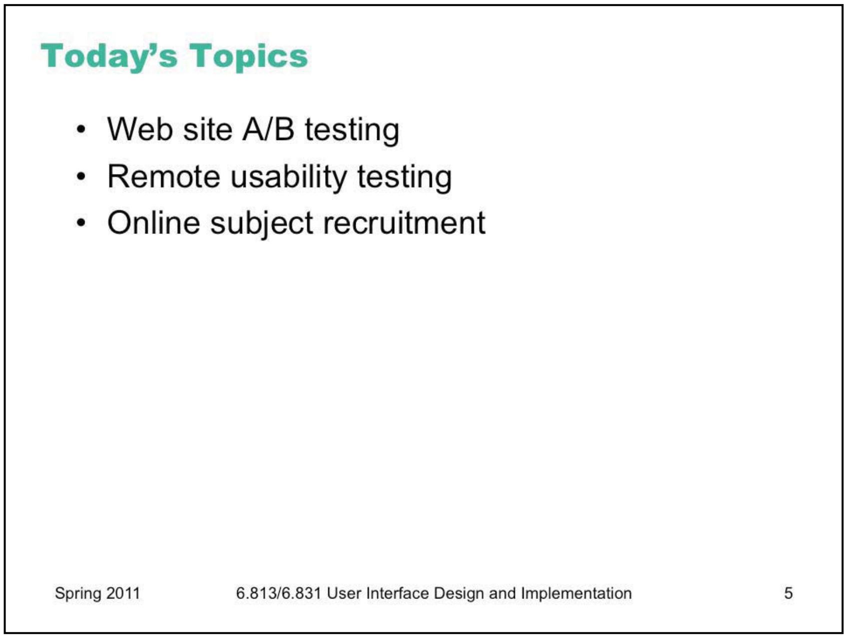
•It’s a modal dialog box, so all this information will go away as soon as the user needs to get to the catalog (**minimize memory load**)

•Using technical terms like V.90 modem (**speak the user’s language**)

•“Please carefully jot down the Model Numbers” (**recognition, not recall**)

•Poor response times: 20-60 second response times (**user control and freedom**), though in fairness this was common for the web at the time, and maybe Acrobat has sufficient progress interfaces to make up for it.

•Misspelling “our catalog” in the first line (**speak the user’s language**)



Today's Topics

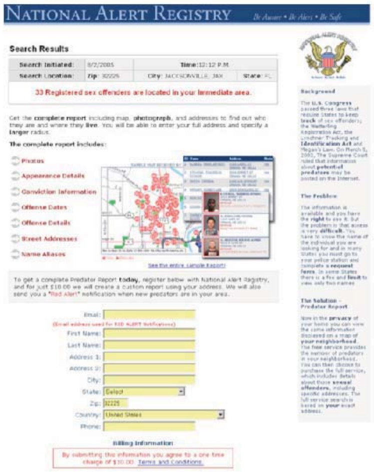
- Web site A/B testing
- Remote usability testing
- Online subject recruitment

Spring 2011 6.813/6.831 User Interface Design and Implementation 5

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.

The content of this lecture is based in part on Kohavi, Henne, Sommerfeld, "Practical Guide to Controlled Experiments on the Web", *KDD 2007*. http://exp-platform.com/Documents/GuideControlledExperiments.pdf

Which Is Better?



NATIONAL ALERT REGISTRY *Be Aware • Be Alert • Be Safe*

Search Results

Search Initiated:	8/2/2005	Time:12:12 P.M.	
Search Location:	Zip: 32225	City: JACKSONVILLE, JAX	State: FL

33 Registered sex offenders are located in your immediate area.

Get the **complete report** including map, **photograph**, and addresses to find out who they are and where they **live**. You will be able to enter your full address and specify a **larger** radius.

The complete report includes:

- Photos
- Appearance Details
- Conviction Information
- Offense Dates
- Offense Details
- Street Addresses
- Name Aliases

See the entire sample Report!

To get a complete Predator Report **today**, register below with National Alert Registry, and for just $10.00 we will create a custom report using your address. We will also send you a "Red Alert" notification when new predators are in your area.

Email:
(Email address used for RED ALERT Notifications)
First Name:
Last Name:
Address 1:
Address 2:
City:
State: Select
Zip: 32225
Country: United States
Phone:

Billing Information

By submitting this information you agree to a one time charge of $10.00. Terms and Conditions.

Background

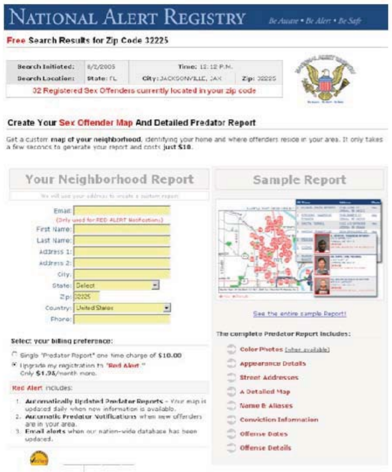
The **U.S. Congress** passed three laws that require States to keep **track** of sex offenders, the Wetterling Registration Act, the Lychner Tracking and **Identification Act** and Megan's Law. On March 5, 2003, The Supreme Court ruled that information about **potential predators** may be posted on the Internet.

The Problem

The information is available and you have the **right** to see it. But the problem is that access is very **difficult**. You have to know the name of the individual you are looking for and in many States you must go to your police station and complete a **request form**. In some States there is a fee and **limit** to view only two names.

The Solution – Predator Report

Now in the **privacy** of your home you can view the same information displayed on a map of **your neighborhood**. The free service provides the number of predators in your neighborhood. You can then choose to purchase the full service, which includes details about those **sexual offenders**, including specific addresses. The full service search is based on **your** exact address.



NATIONAL ALERT REGISTRY *Be Aware • Be Alert • Be Safe*

Free Search Results for Zip Code 32225

Search Initiated:	8/2/2005	Time: 12:12 P.M.	
Search Location:	State: FL	City: JACKSONVILLE, JAX	Zip: 32225

32 Registered Sex Offenders currently located in your zip code

Create Your Sex Offender Map And Detailed Predator Report

Get a custom **map of your neighborhood**, identifying your home and where offenders reside in your area. It only takes a few seconds to generate your report and costs **just $10**.

Your Neighborhood Report

We will use your address to create a custom report

Email:
(Only used for RED ALERT Notifications)
First Name:
Last Name:
Address 1:
Address 2:
City:
State: Select
Zip: 32225
Country: United States
Phone:

Select your billing preference:

○ Single "Predator Report" one time charge of **$10.00**
● Upgrade my registration to "Red Alert." Only **$1.95**/month more.

Red Alert includes:

1. **Automatically Updated Predator Reports** – Your map is updated daily when new information is available.
2. **Automatic Predator Notifications** when new offenders are in your area.
3. **Email alerts** when our nation-wide database has been updated.

Sample Report

See the entire sample Report!

The complete Predator Report includes:

- Color Photos (when available)
- Appearance Details
- Street Addresses
- A Detailed Map
- Name & Aliases
- Conviction Information
- Offense Dates
- Offense Details

© National Alert Registry. All rights reserved.

Spring 2011 6.813/6.831 User Interface Design and Implementation 6

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.

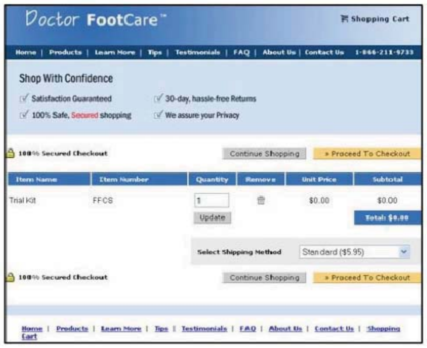
We could look closely at these examples and pick them apart with respect to usability principles (visibility, learnability, efficiency, etc.), and the designers were doubtlessly thinking about principles and justifications for the design decisions they made. But at the end of the day, which design is 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**.

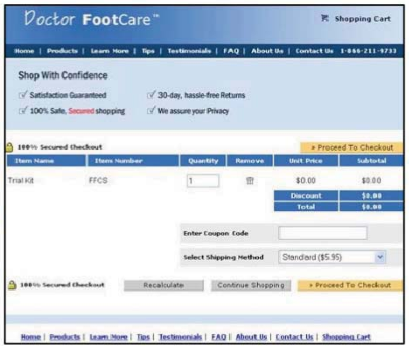
Source: http://www.alistapart.com/articles/designcancripple

Which is Better?


Doctor FootCare™
Shopping Cart
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us 1-866-211-9733
Shop With Confidence
Satisfaction Guaranteed
30-day, hassle-free Returns
100% Safe, Secured shopping
We assure your Privacy
100% Secured Checkout
Continue Shopping
» Proceed To Checkout
Update
100% Secured Checkout
Continue Shopping
» Proceed To Checkout
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping Cart

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1	🗑	$0.00	$0.00
					Total: $0.00

Select Shipping Method: Standard ($5.95)

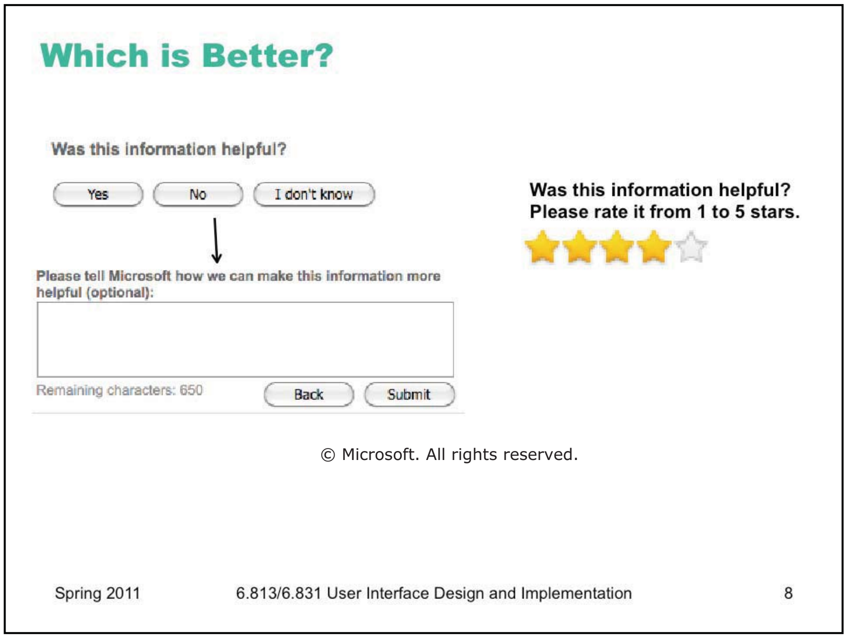

Doctor FootCare™
Shopping Cart
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us 1-866-211-9733
Shop With Confidence
Satisfaction Guaranteed
30-day, hassle-free Returns
100% Safe, Secured shopping
We assure your Privacy
100% Secured Checkout
» Proceed To Checkout
100% Secured Checkout
Recalculate
Continue Shopping
» Proceed To Checkout
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping Cart

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1	🗑	$0.00	$0.00
				Discount	$0.00
				Total	$0.00

Enter Coupon Code:

Select Shipping Method: Standard ($5.95)

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.



Which is Better?

Was this information helpful?

Please tell Microsoft how we can make this information more helpful (optional):

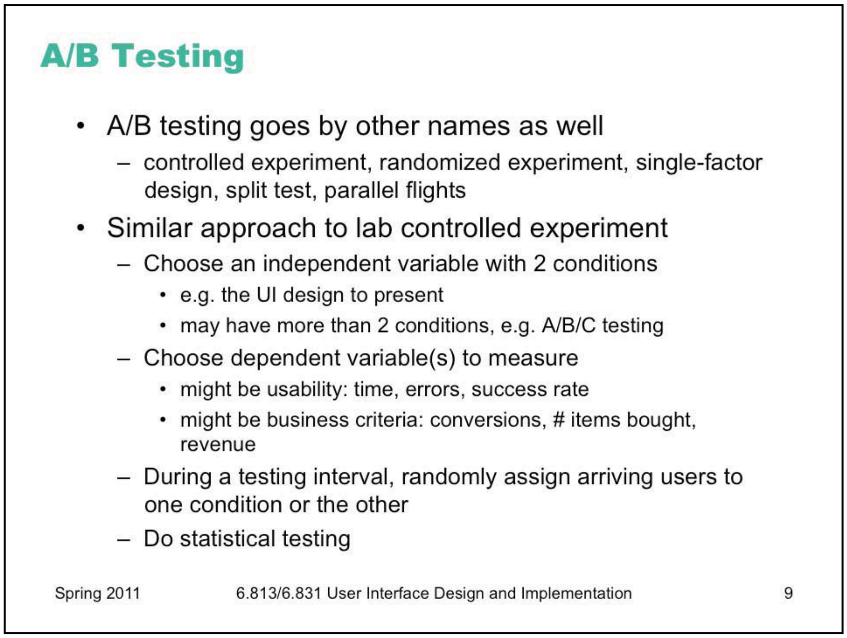
Remaining characters: 650

Was this information helpful?
Please rate it from 1 to 5 stars.

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.



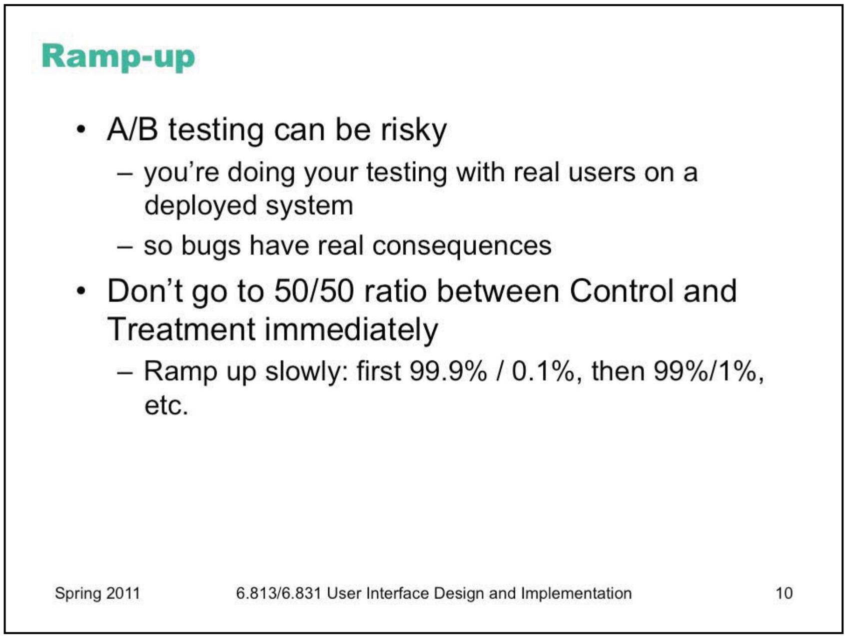
A/B Testing

- A/B testing goes by other names as well
  - controlled experiment, randomized experiment, single-factor design, split test, parallel flights
- Similar approach to lab controlled experiment
  - Choose an independent variable with 2 conditions
    - e.g. the UI design to present
    - may have more than 2 conditions, e.g. A/B/C testing
  - Choose dependent variable(s) to measure
    - might be usability: time, errors, success rate
    - might be business criteria: conversions, # items bought, revenue
  - During a testing interval, randomly assign arriving users to one condition or the other
  - Do statistical testing

Spring 2011 6.813/6.831 User Interface Design and Implementation 9

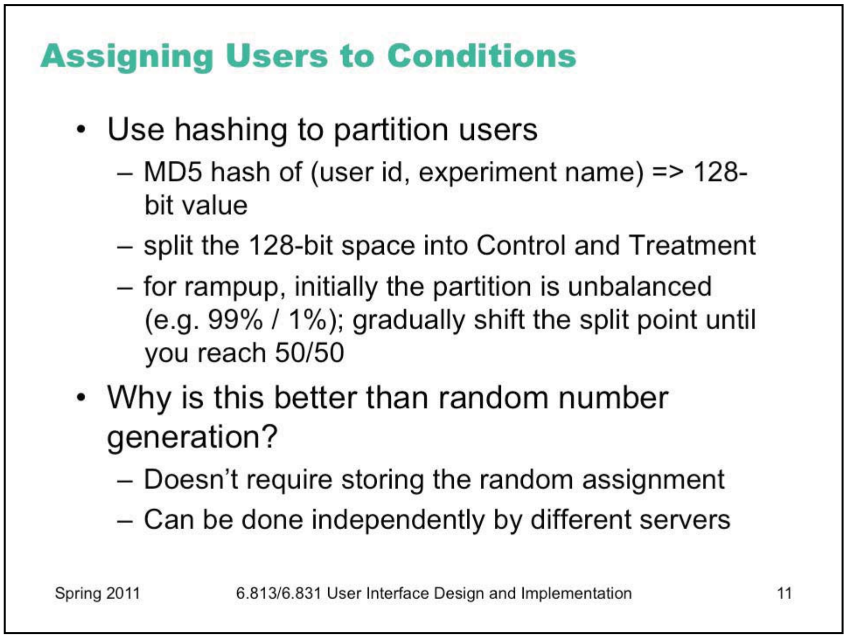
The term "A/B testing" actually comes from marketing. Other fields have other names for the idea – in the context of usability studies in the lab, we've been calling them controlled experiments. The setup is basically the same: you choose an independent variable (like the UI design) with at least two 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.



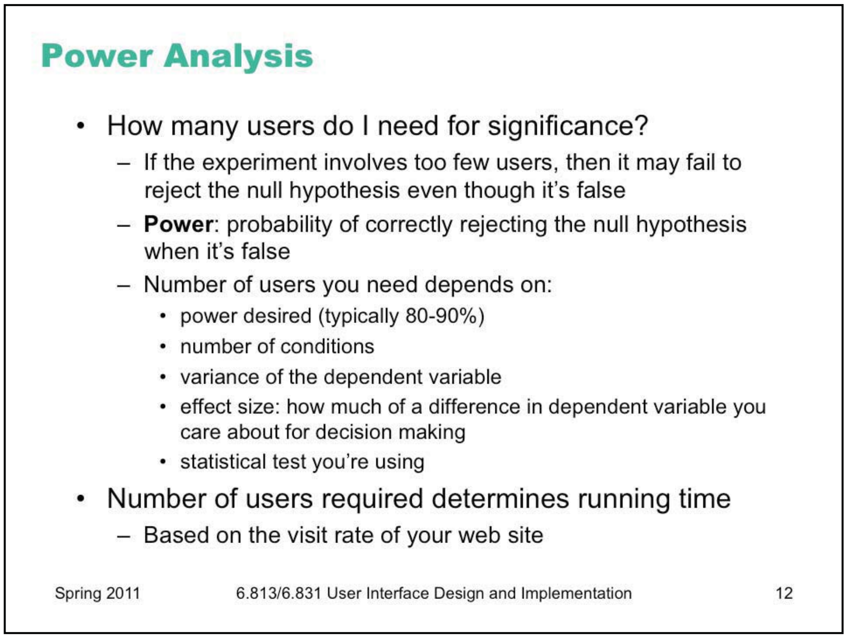
Ramp-up

- 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.



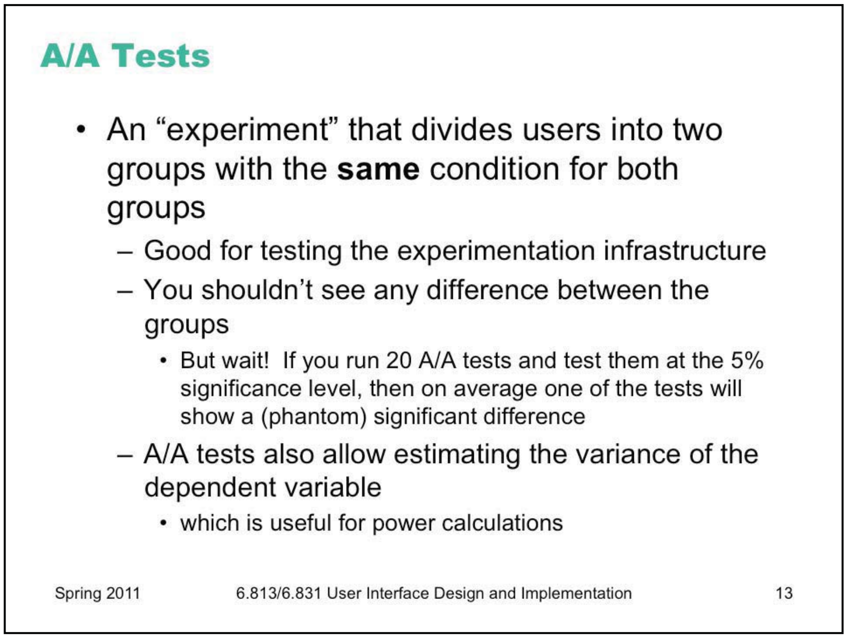
Assigning Users to Conditions

- 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



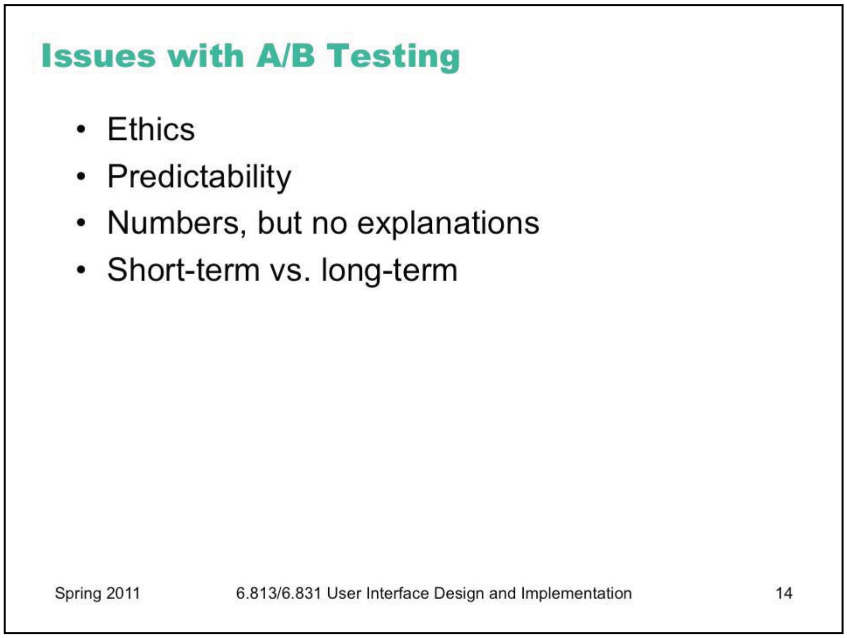
Power Analysis

- 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



Issues with A/B Testing

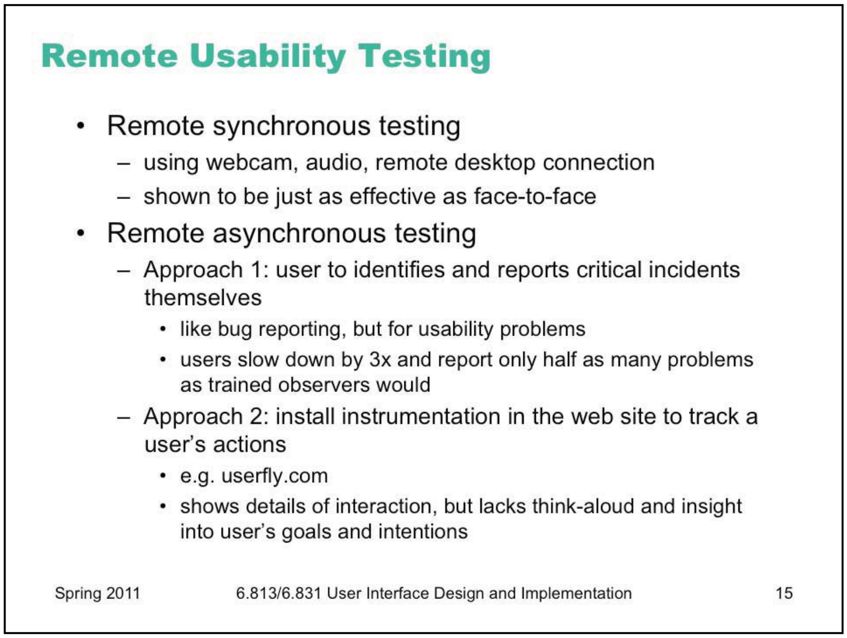
- Ethics
- Predictability
- Numbers, but no explanations
- Short-term vs. long-term

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.

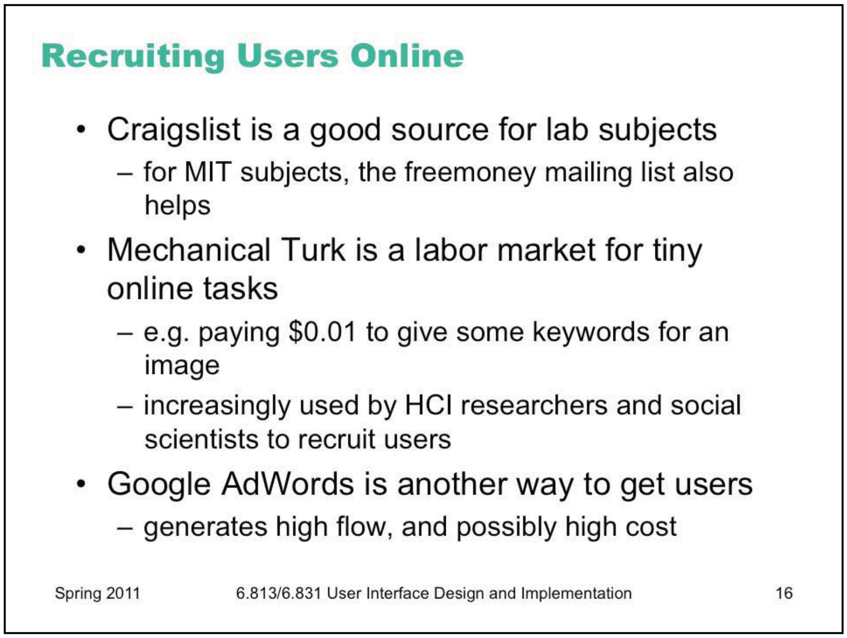
Short-term vs. long-term: a typical A/B test runs only for days or weeks, while the real effect of a new design might be seen only over a long term, as users learn how to use it well. But it's worth noting that even days or weeks is a longer term than a typical lab-based user study, which might last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

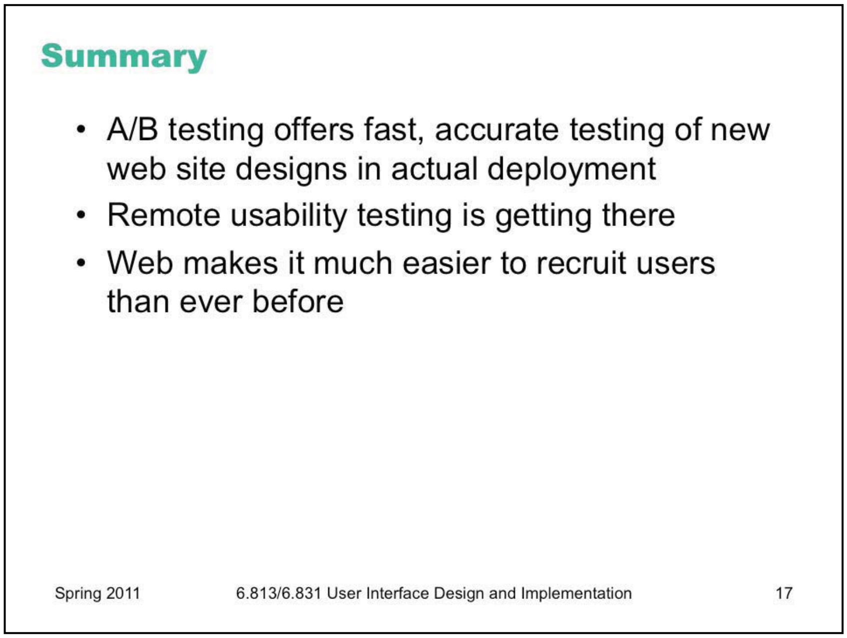
See Andreason et al., "What Happened to Remote Usability Testing? An Empirical Study of Three Methods", CHI 2007.



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

See Kittur, Chi, Suh, “Crowdsourcing user studies with Mechanical Turk”, CHI 2008.



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

6.831 / 6.813 User Interface Design and Implementation
Spring 2011