Evaluation studies: From controlled to natural settings

Lecture 14

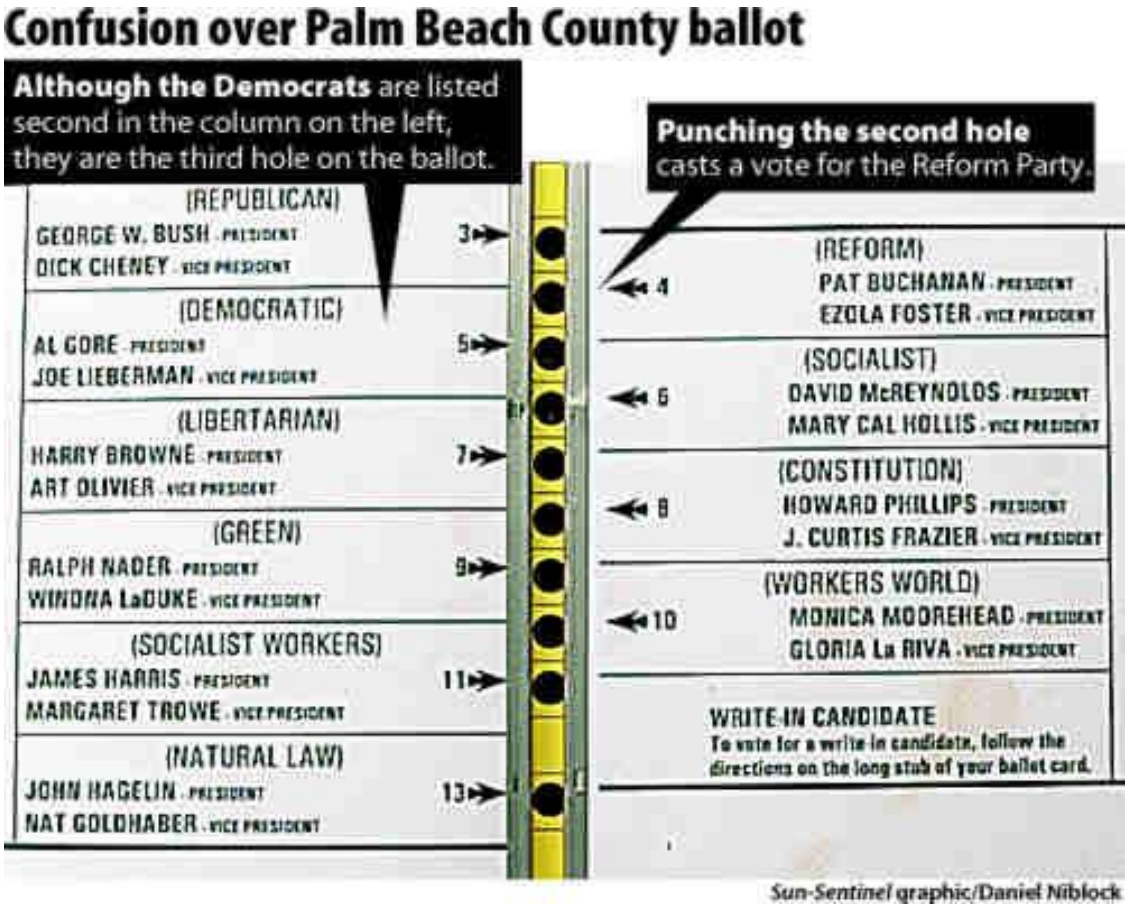
Kristina Lapin
Vilnius University

Slides adapted from

The aims:

- Explain how to do usability testing
- Outline the basics of experimental design
- Describe how to do field studies

Motivation



http://www.asktog.com/columns/042ButterflyBallot.html

Testing with users

- Why?
  - Early find problems
    - Before coding
- When?
  - Specyfying requirements
  - designing
  - testing
  - Before deployment

How many participants?

- 5 participants reveal about 80% defects
- But:
  - Test usefulness depends on participants and system
- Simple answer
  - 5 participant reveal suffucient number of defects

(Virzi, 1992, Nielsen, Landauer, 1993)

Testing process

- Planing
- Involving participants
- Defining tasks
- Pilot testing
- Testing
- Analysing results
- Writing report

Usability testing

- Involves recording performance of typical users doing typical tasks.
- Controlled settings.
- Users are observed and timed.
- Data is recorded on video & key presses are logged.
- The data is used to calculate performance times, and to identify & explain errors.
- User satisfaction is evaluated using questionnaires & interviews.
- Field observations may be used to provide contextual understanding.

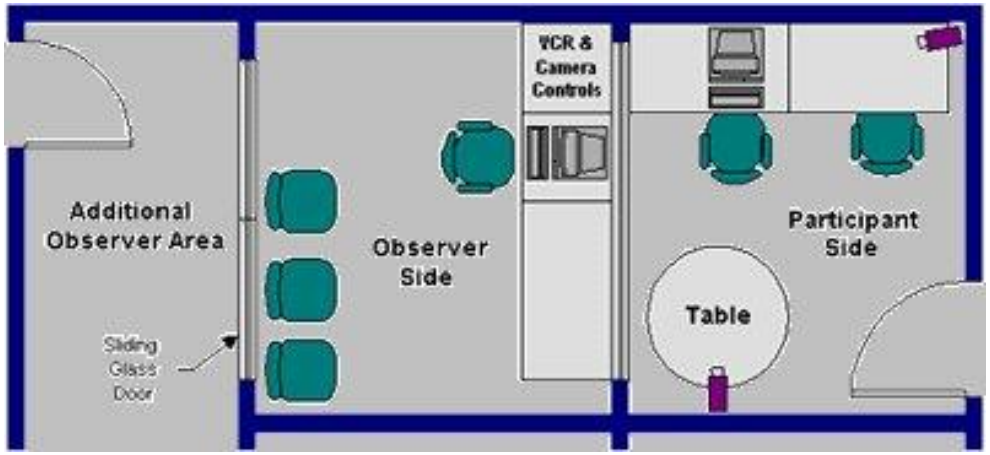
Usability testing

- Goals & questions focus on
  - how well users perform tasks with the product.
  - Comparison of products or prototypes.
- Focus is on time to complete task & number & type of errors.
- Data collected by video & interaction logging.
- Testing is central.
- User satisfaction questionnaires & interviews provide data about users' opinions.

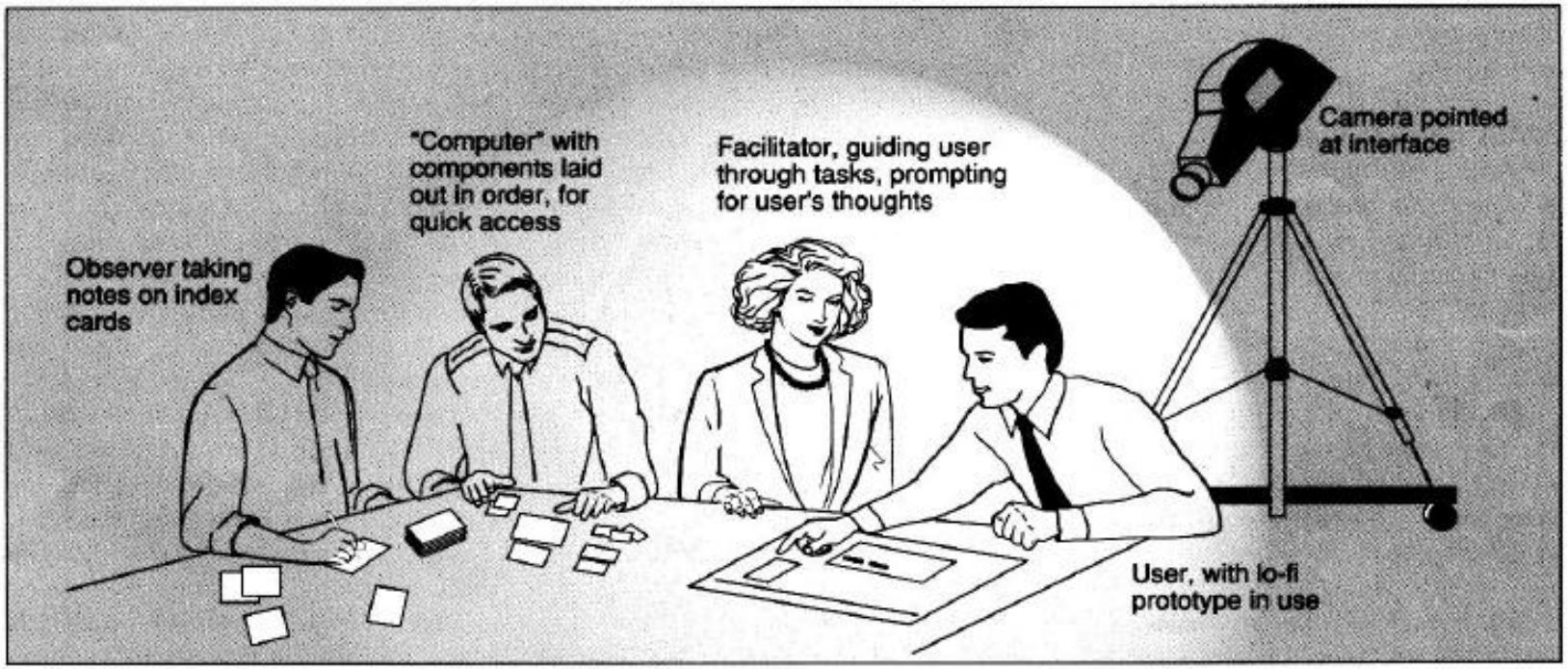
Usability lab with observers watching a user & assistant

Usability lab

- 1-3 video cameras, microphons
- Camera remote control



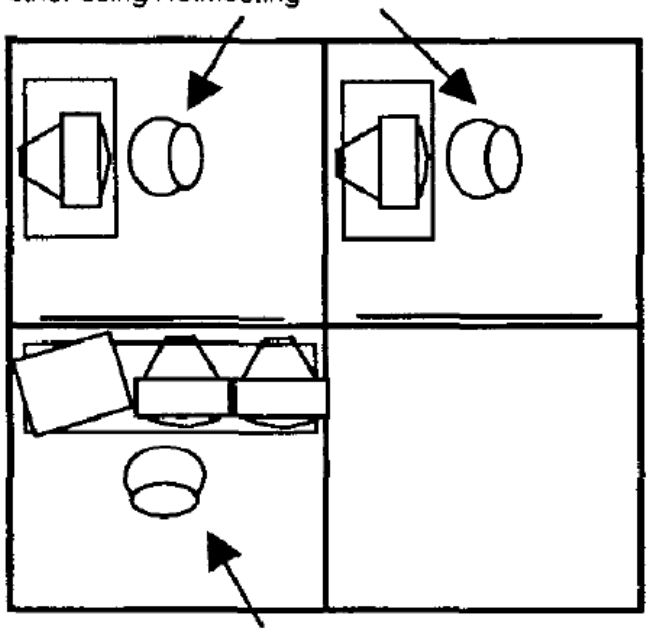
Testing a paper prototype



http://www.iua.upf.es/~jblat/material/hci/2008/paper_prototyping/p21-rettig.pdf

Use case: the testing of NetMeeting, an early videoconferencing product

Evaluation Participants communicating with each other using NetMeeting



Usability engineer uses another PC to become the third participant.

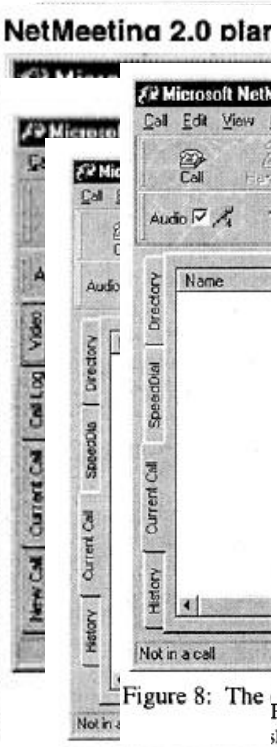


Figure 8: The

Figure 5

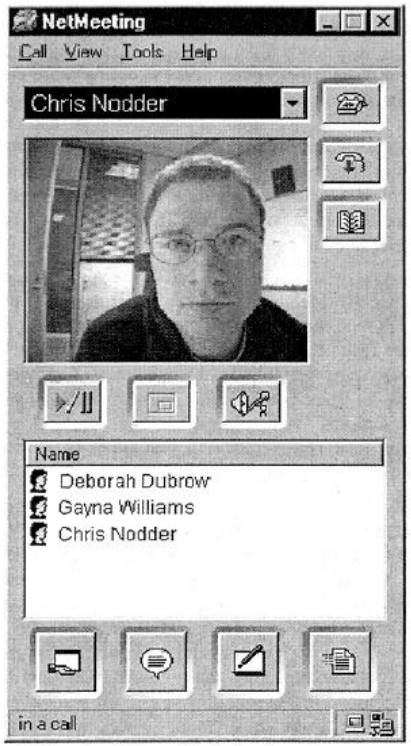


Figure 9: Version 3 Beta 1default view. A compact view, showing just the picture or just the sharing controls is also available.

Chris Nodder, Gayna Williams, Deborah Dubrow (1999) Evaluating the usability of an evolving collaborative product - changes in user type, tasks and evaluation methods over time. Proceedings of the international ACM SIGGROUP conference on Supporting group work, 1999. http://www.cis.gvsu.edu/~laa/CS623/p150-nodder.pdf

Use case: the testing of NetMeeting, an early videoconferencing product

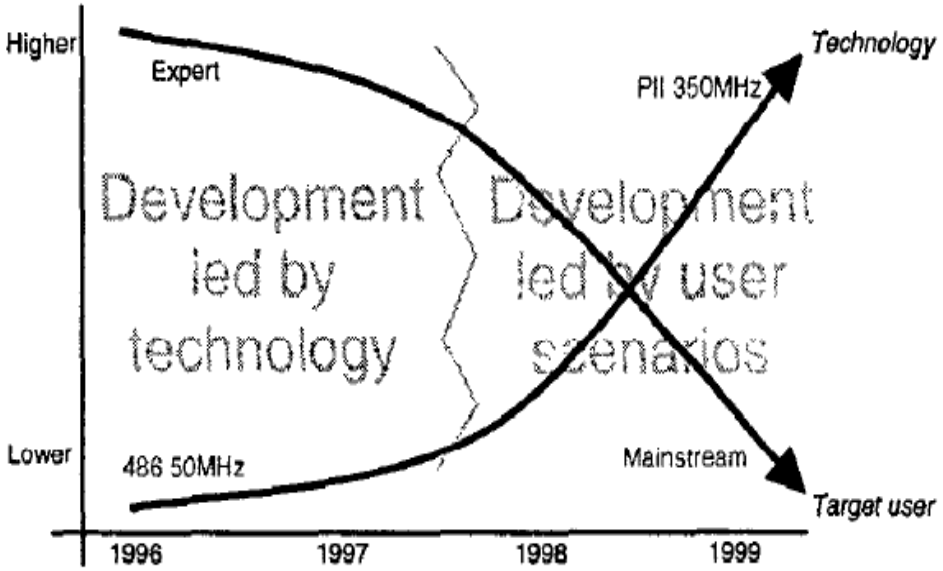


Figure 10: Change in focus from technology to user scenarios as NetMeeting matured

Chris Nodder, Gayna Williams, Deborah Dubrow (1999) Evaluating the usability of an evolving collaborative product - changes in user type, tasks and evaluation methods over time. Proceedings of the international ACM SIGGROUP conference on Supporting group work, 1999. http://www.cis.gvsu.edu/~tao/CS623/p150-nodder.pdf

Portable equipment for use in the field

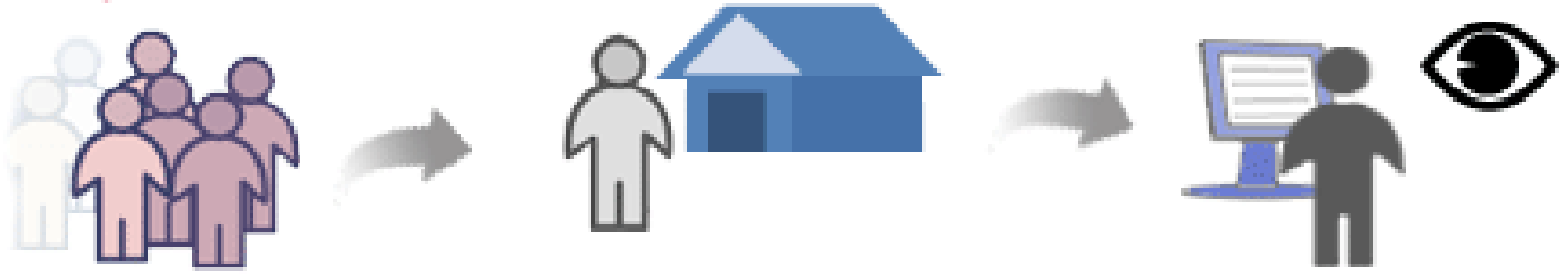
Tracksys portable lab include: camera with direct plug to PC, software GoToMeeting, remote control system, new eye-tracking devices

Remote Usability Testing

A selected group of panelists are invited to participate

...They are asked to evaluate the web from their natural context, using Internet Explorer

...A robot (UZ Bar) guides the users and monitors their behavior



The users are asked to complete certain tasks and answer questions

The UZ Platform gatheres and saves the data in real-time

Servidores UserZoom

The data is analysed and a final report is prepared

Remote usability testing

- Handreds of users can be tested
- Participation in the natual context from geographically spread locations
- No human moderation needed

http://www.slideshare.net/UserZoom/case-study-lab-online-usability-testing-4041695?from=ss_embed

Testing conditions

- Usability lab or other controlled space.
- Emphasis on:
  - selecting representative users;
  - developing representative tasks.
- 5-10 users typically selected.
- Tasks usually last no more than 30 minutes.
- The test conditions should be the same for every participant.
- Informed consent form explains procedures and deals with ethical issues.

Measures (metrics)

- Time to complete a task.
- Time to complete a task after a specified. time away from the product.
- Number and type of errors per task.
- Number of errors per unit of time.
- Number of navigations to online help or manuals.
- Number of users making a particular error.
- Number of users completing task successfully.

Usability engineering orientation

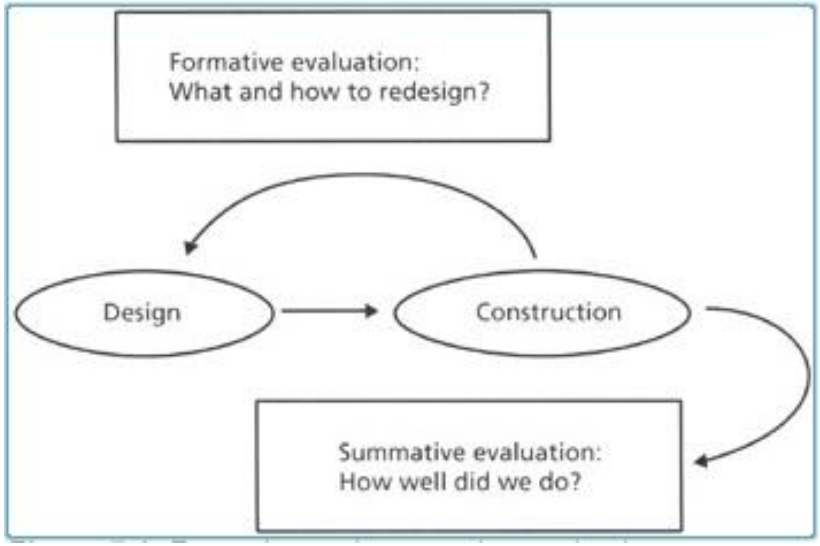
- Aims
  - to improve with each version.
  - Measure current level of performance.
- Define
  - Minimum acceptable level of performance.
  - Target level of performance.

How many participants is enough for user testing?

- The number is a practical issue.
- Depends on:
  - schedule for testing;
  - availability of participants;
  - cost of running tests.
- Typically 5-10 participants.
- Some experts argue that testing should continue until no new insights are gained.

User Testing in the Design Process

- Empirical evaluation can happen at every stage
- Formative evaluation
  - Happens throughout the design process
  - Can evaluate scenarios, sketches, models, prototypes
- Summative evaluation
  - Typically happens at the end
  - Assesses system and interface design quality, i.e., how well have we done?



User Testing

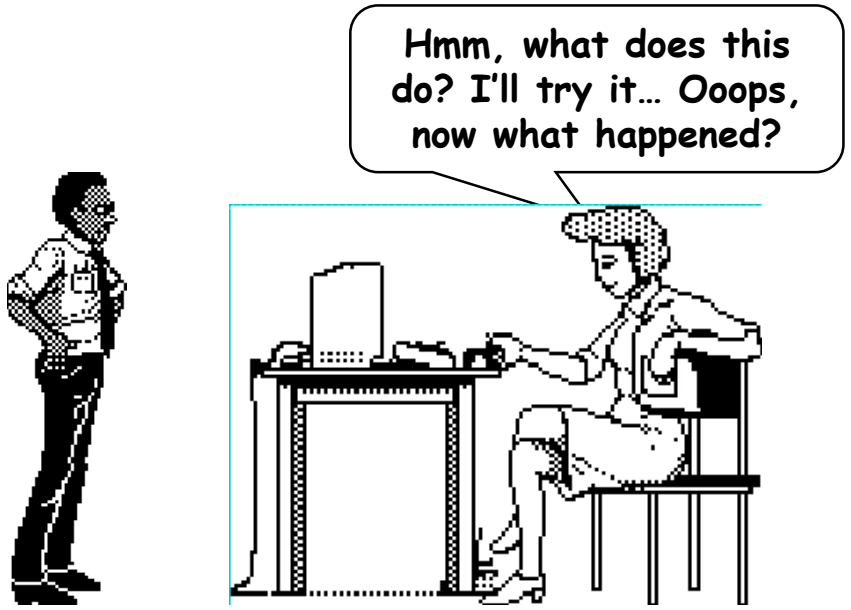
- Methods
  - Design and implement scenario or prototype
  - Record user behaviour
    - Typical usage, or critical incidents
    - Keystroke recording
    - Thinking aloud protocols
    - Videotape protocols
  - Interviews for subjective impressions
  - Analyze user behaviour
- Roles
  - Understanding user methods
  - Understanding user problems
  - Discovering user thought processes

Tipiniai tyrimo tikslai

- Laikas
  - įvykdyti užduotį,
  - rasti tinklapį
- Klaidų skaičius
  - per užduotį
  - per laiko tarpą
  - tam tikro tipo
- Kreipinių į pagalbą skaičius
- Dalyvių, sėkmingai baigusių užduotį, skaičius
- Priežastys, dėl kurių dalyvis nebaigė užduoties.

Think aloud protocol

- In this approach the user says out what she is thinking while she is carrying out a task or doing some problem solving. User's thinking is recorded as an audio record.
- Using this protocol we collect the reactions of the users.
- This is quite helpful because many aspects of human behavior and mind is not predictable by engineering models.



User Testing

- Carrying out the study
  1. Let users know that complete anonymity will be preserved
  2. Let them know that they may quit at any time
  3. Stress that the system is being tested, not the participant
  4. Indicate that you are only interested in their thoughts relevant to the system
  5. Demonstrate the thinking-aloud method by acting it out for a simple task, e.g., figuring out how to load a stapler

User Testing

- Carrying out the study
  6. Hand out instructions for each part of the study individually, not all at once
  7. Maintain a relaxed environment free of interruptions
  8. Occasionally encourage users to talk if they grow silent
  9. If users ask questions, try to get them to talk (e.g., “What do you think is going on?” and follow predefined rules on when to help or interrupt to help.
  10. Debrief each user after the experiment

User Testing

- Improving the study
  - The pilot study should "debug" the study. This minimize changes during the study, allowing quantitative data analysis. But improvements may be warranted.
  - Experimenters' role can be improved
  - Tasks given to participant can be improved
  - Written materials can be improved

Usability Test Documents

1. A usability test informed consent
   - Informs the user about the test and provides formal agreement by the user to participate
2. A usability test script
   - Details the user actions
3. A pre-test questionnaire
   - User age, gender, occupation, used technologies
4. A post-test questionnaire
   - Asks the participants to describe their experience

Usability test report

- Executive summary
- Introduction
- Participants
- Methods
- Findings and recommendations
- Conclusions
- Appendices

EXAMPLE OF THE USABILITY TESTING

Use case: name 3 features for each that can be tested by usability testing

iPad

Testing goals

- Ar user expectations different for the iPad compared with iPhone?
  - Previous study of the iPhone:
    - people preferred using apps than browsing the web
    - because the latter is slow and cumbersome
- Whether it is worth developing specific websites for the iPad (like for smartphones)?
  - Or the desktop versions are acceptable?

http://www.nngroup.com/reports/mobile/ipad/

Participants

- Seven participants:
  - All experienced iPhone users who
    - had owned iPhone for at least 3 months
    - had used a variety of apps.
  - Age: 20-60
  - Occupations: food server, legal, medical staff, retired driver, homemaker, accounter
  - 3 males, 4 females

http://www.nngroup.com/reports/mobile/ipad/

Tasks

- In the beginning: Ad-hoc tasks.
  - Examples of used apps:

APP	TASK
Adobe Idea	Draw a sketch of your apartment.
Amazon Mobile	Find a birthday gift for yourself.
Bloomberg	How do you display your favorite news topics on the first page?
The Daily	Find a story of interest and make sure you can get back to it later.
Fandango	Find a movie you may want to watch during the weekend and buy tickets for it.
Indian Vegetarian Restaurant	Look for a vegetarian restaurant around this area.

http://www.nngroup.com/reports/mobile/ipad/

Specific tasks

APP OR WEBSITE	TASK
ABC News	Check the latest news.
Amazon Windowshop (amazon.com)	Look for a birthday gift for yourself.
Amazon Windowshop (amazon.com)	Look for a flexible iPad keyboard.
BigOven	Find a recipe for lamb roast.
Bing	Check the latest world news.
Bing	You're going to the movies on Friday night. Find a movie to watch.
The Daily	Find the latest news about the earthquake in Japan.
Flipboard	Check the latest news. Set up the app to show the news topics that interest you.
Fortune	Find an article about the President's plan to deal with the housing crisis.
Fortune	Figure out what makes the largest part of the cost of an airplane ticket.

http://www.nngroup.com/reports/mobile/ipad/

The equipment

Mobile usability kit

Procedure

- Camera recorded interactions and gestures using iPad
- Webcam – xpressions of participants' faces and think-aloud commentary.
- Observers watched the video
  - rather than observing directly

http://www.nngroup.com/reports/mobile/ipad/

The findings

- The participants were able to interact with websites on iPad but it was not optimal
  - Links too small to tap on reliably
  - The fonts sometimes dificult to read
- Usability problems were classified to interaction design principles:
  - Mental models, navigation, the quality of images, touschscreen problems, lack of affordances, getting lost in the application, working memory, and the received feedback.

http://www.nngroup.com/reports/mobile/ipad/

Users can select the contents page from the contents carousel.

Clicking on the Contents button opens the contents carousel.

Contents

TIME

Images link to the corresponding article.

To Our Readers
For the first time in our history, TIME will deliver the magazine's content in a radically different way: as a self-contained application you can download to the iPad

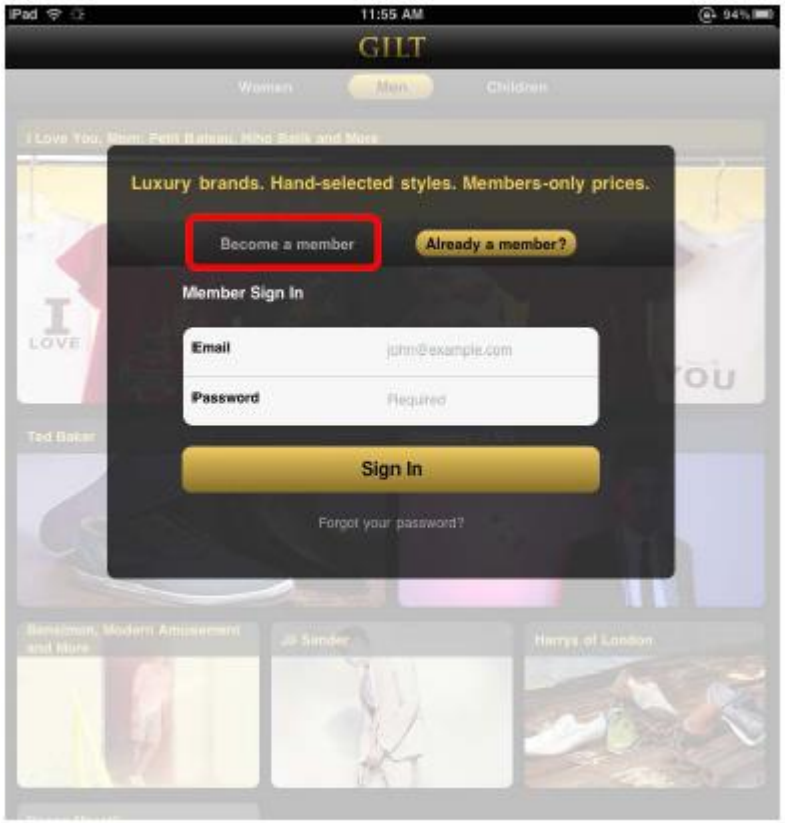
The Moment
Suicide bombings killed at least 39 people in the worst terrorist attacks to hit Moscow in six years. Can Russia suppress the new threats?
by Simon Shuster

TimeFrames
Imelda Marcos plants a kiss on the glass coffin containing the preserved body of her late husband Ferdinand, who died in 1989. Plus, more Pictures of the Week

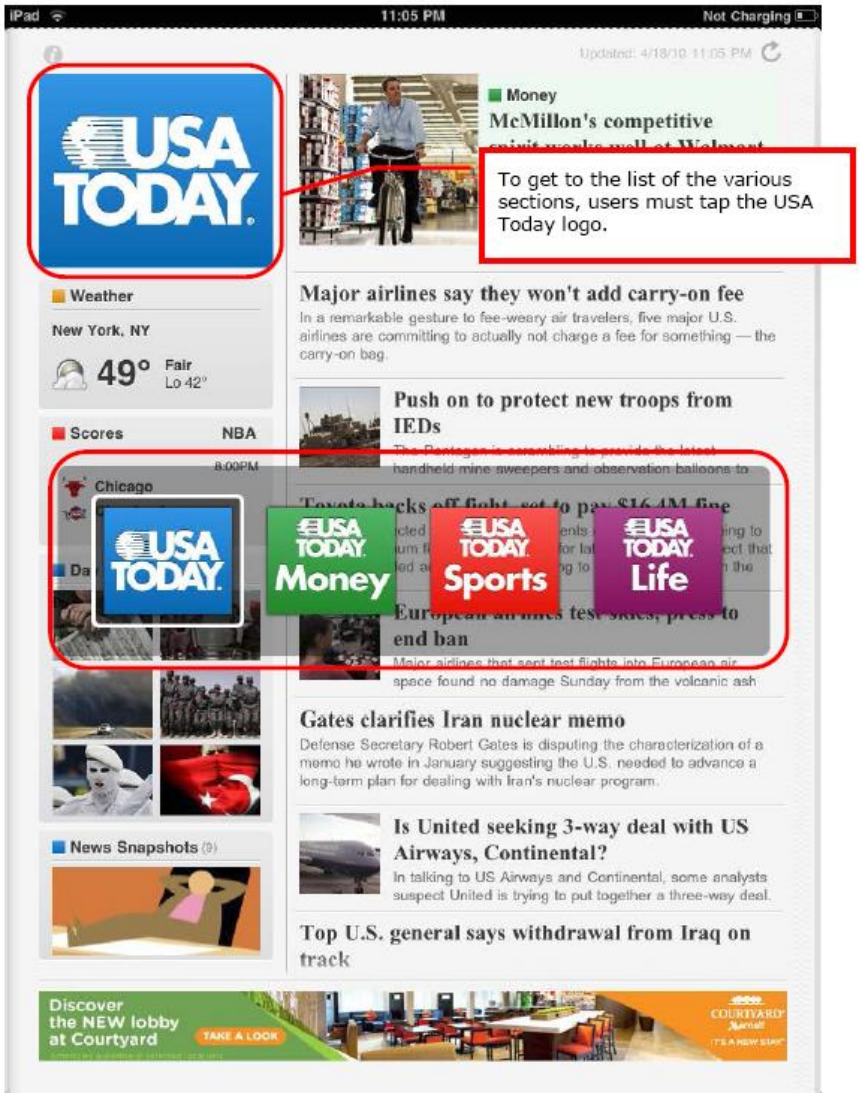
BRIEFING

The World
Burmese opposition boycotts elections; teen bullies indicted in Massachusetts; Berlusconi passes power test; an epic proton smash; an apology from Serbia; an island vanishes

Lack of Affordances: Where Can I Tap?



Getting Lost in an Application

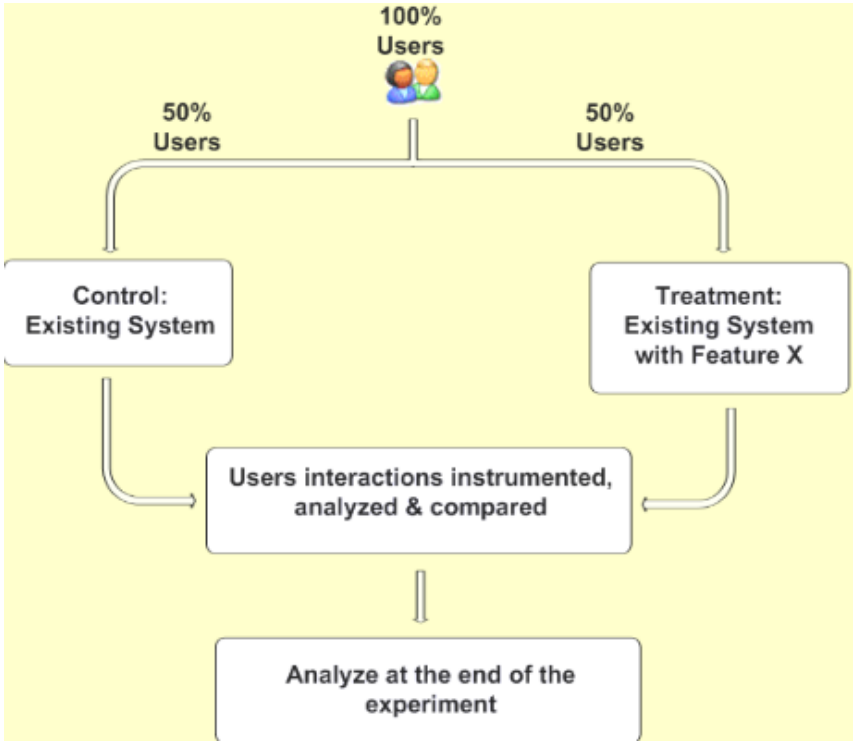


CONTROLLED EXPERIMENTS

Controlled experiments

- Predict the relationship between two or more variables.
  - A/B testing
- Independent variable is manipulated by the researcher.
- Dependent variable depends on the independent variable.
- Typical experimental designs have one or two independent variable.
- Validated statistically & replicable.

A/B testing



- Participants are divided across the conditions
- Compare results

Practical Guide to Controlled Experiments on the Web, Ron Kohavi, Randal Henne, Dan Sommerfield, *KDD 2007: ACM Conference on Knowledge Discovery and Data Mining.*

Hypotheses testing

- A hypothesis tests the effect of the independent variable on the dependent variable
  - A null hypothesis
    - No difference between dependent variables
  - Alternative hypothesis
    - There is a difference

Experimental designs

- Different participants - single group of participants is allocated randomly to the experimental conditions.
- Same participants - all participants appear in both conditions.
- Matched participants - participants are matched in pairs, e.g., based on expertise, gender, etc.

Example: structure in web page design

- The goals of experiment was to find the optimal depth versus breadth structure of hyperlinks
  - Condition 1: 8 x 8 x 8
  - Condition 2: 16 x 32
  - Condition 3: 32 x 16
  - A same-participant experiment, random tasks
- Results
  - C1: reaction time = 58 sec., SD = 23
  - C2: reaction time = 36 sec, SD=16
  - C3: reaction time = 46 sec, SD=26
- Conclusion: breadth is preferable to depth.

Larson, Czerwinski 1998

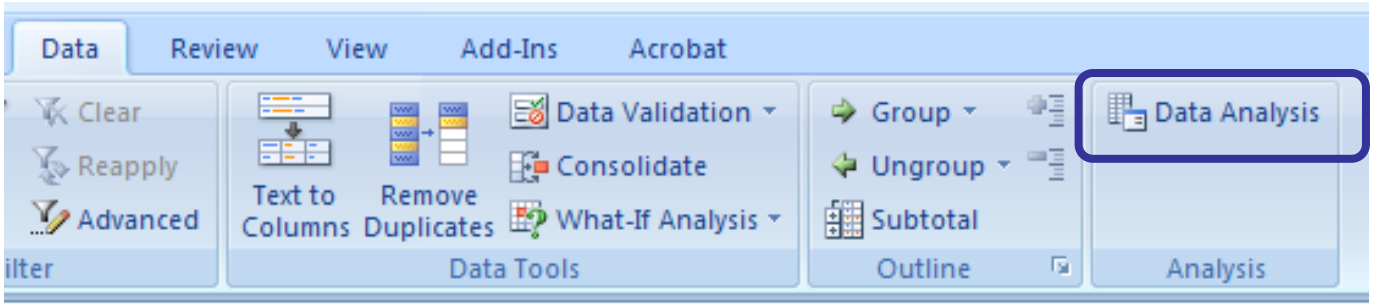
Different, same, matched participant design

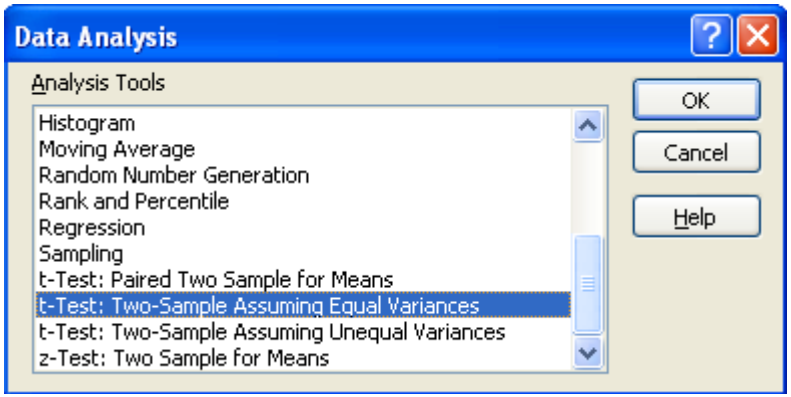
Design	Advantages	Disadvantages
Different	No order effects	Many subjects & individual differences a problem
Same	Few individuals, no individual differences	Counter-balancing needed because of ordering effects
Matched	Same as different participants but individual differences reduced	Cannot be sure of perfect matching on all differences

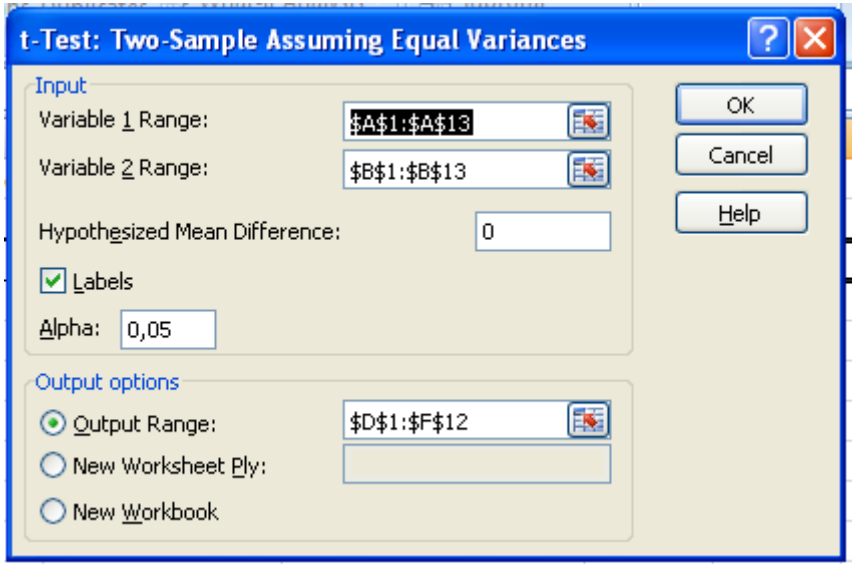
Statistics: t-tests

- The measure results are used to compute the means and standard deviations (SD)
  - SD – statistical measure of the spread or variabilityarount the mean.
- T-test tests a significance of the diference beteen the means for the two condidition

T-test (MS Excel)







T-test (MS Excel)

A	B	C	D	E	F
Expert_time	Novice_time		t-Test: Two-Sample Assuming Equal Variances		
34	46				
33	48			*Expert_time*	*Novice_time*
28	53		Mean	35,1	49,4
44	66		Variance	126,4	229,0
46	67		Observations	12	12
21	35		Pooled Variance	177,7	
22	39		Hypothesized Mean Difference	0	
53	21		df	22	
22	34		t Stat	-2,63	
29	55		P(T<=t) one-tail	0,0076	
39	59		t Critical one-tail	1,717	
50	70		P(T<=t) two-tail	0,0152	
			t Critical two-tail	2,0739	

Null hypothesis

Experts are faster than novices

Usability testing & research

Usability testing

- Improve products
- Few participants
- Results inform design
- Usually not completely replicable
- Conditions controlled as much as possible
- Procedure planned
- Results reported to developers

Experiments for research

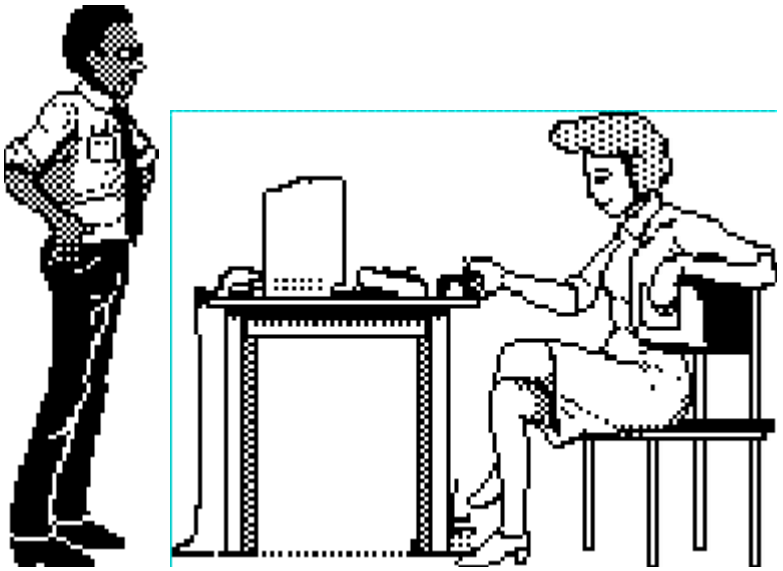
- Discover knowledge
- Many participants
- Results validated statistically
- Must be replicable
- Strongly controlled conditions
- Experimental design
- Scientific report to scientific community

Field studies

- Field studies are done in natural settings.
  - "in the wild" is a term for prototypes being used freely in natural settings.
- Aim to understand what users do naturally and how technology impacts them.
- Field studies are used in product design to:
  - identify opportunities for new technology;
  - determine design requirements;
  - decide how best to introduce new technology;
  - evaluate technology in use.

Observing

- Natural setting
  - More time to organise and conduct
  - More efforts to analyse the results
- User performs the task
  - Unclear reasons of user bahavior



Data collection & analysis

- Observation & interviews
  - Notes, pictures, recordings
  - Video
  - Logging
- Analyzes
  - Categorized
  - Categories can be provided by theory
    - Grounded theory
    - Activity theory

Data presentation

- The aim is to show how the products are being appropriated and integrated into their surroundings.
- Typical presentation forms include: vignettes, excerpts, critical incidents, patterns, and narratives.

Key points

- Usability testing is done in controlled conditions.
- Usability testing is an adapted form of experimentation.
- Experiments aim to test hypotheses by manipulating certain variables while keeping others constant.
- The experimenter controls the independent variable(s) but not the dependent variable(s).
- There are three types of experimental design: different-participants, same- participants, & matched participants.
- Field studies are done in natural environments.
- "In the wild" is a recent term for studies in which a prototype is freely used in a natural setting.
- Typically observation and interviews are used to collect field studies data.
- Data is usually presented as anecdotes, excerpts, critical incidents, patterns and narratives.

References

- Rogers, Sharp, Preece (2015). Interaction design: Beyond Human Computer Interaction. Wiley
- Nielsen Norman Group Reports. Usability of iPad Apps and Websites: 2 Reports With Research Findings
- Larson, K., M. Czerwinski (1998) Web page design: implications of memory, structure and scent for information retrieval. In *Proceedings of CHI'98*, pp. 25-32
- S. Consolvo, D. W. McDonald, T. Toscos, M. Chen, J.E. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith & J. A. Landay. "Activity Sensing in the Wild: A Field Trial of UbiFit Garden," Proceedings of the Conference on Human Factors in Computing Systems: CHI '08, Florence, Italy, (2008), pp.1797-806.