USABILITY TESTING

15 Oct 2014

ERIC PAULOS
www.paulos.net

UNIVERSITY OF CALIFORNIA Berkeley

ANNOUNCEMENTS

PROG 02 – Due Friday (Midnight) :: A month since assigned…

DESIGN 04 Due Friday – Your project! Be bold!

DESIGN 05 Contextual Inquiry, Task Analysis, Competitive Analysis – Due 30 Oct – Plan ahead (after midterm)

Midterm review in Section

No AM Class next Thur

Midterm next Thur 1:10 – 2:30

MIDTERM ON 22 OCT

In class – Actually in Sibley Auditorium

Watch Piazza for details

80 minutes

Closed book & notes

Review on Friday 16 Oct in Section

If you are registered with the DSP office and have special needs, you should received email from us about exam accommodations.

MIDTERM ON 22 OCT

All lecture material, slides, and readings.

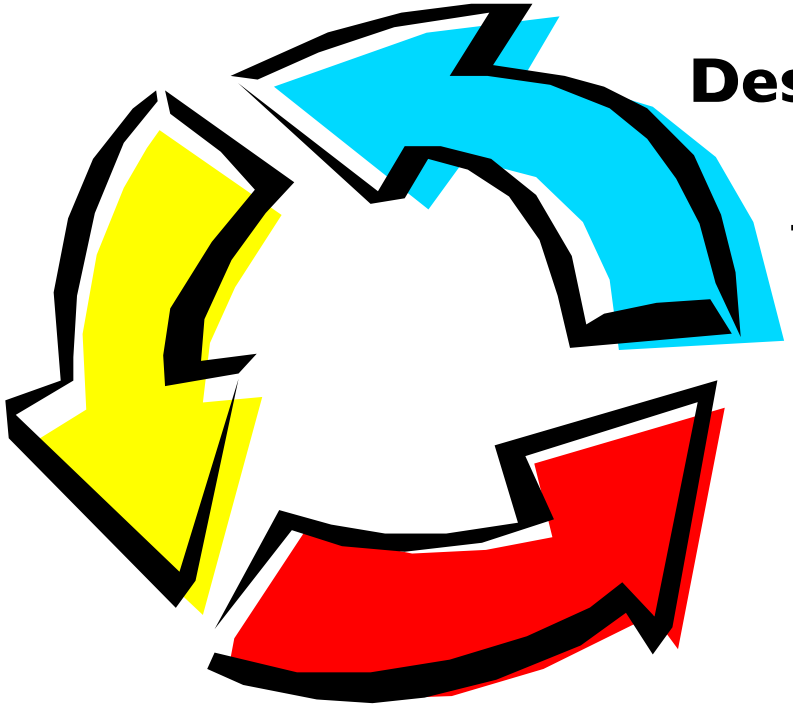
Short answer
Multiple Choice
True / False
Longer descriptions for some

No coding questions

If you find a question ambiguous, document the ambiguity.
Indicate the way you interpreted the question in a set of separate sentences next to the question. The questions on the exam are not intended to be ambiguous, but sometimes another meaning is interpreted by the examinee that we did not take into consideration.

USABILITY TESTING METHODS

ITERATIVE DESIGN



Design

Brainstorming
Task analysis
Contextual inquiry

Prototype

Low-fi, paper

Evaluate

Low-fi testing,
Qualitative eval
Quantitative eval

GENRES OF ASSESSMENT

Automated Usability measures computed by software

Inspection Based on skills, and experience of evaluators

Formal Models and formulas to calculate measures

Empirical Usability assessed by testing with real users

EMPIRICAL TESTING IS COSTLY

User studies are very expensive – you need to schedule (and normally pay) many subjects.

User studies may take many hours of the evaluation team's time.

A user test can easily cost $10k's

“DISCOUNT USABILITY” TECHNIQUES

Cheap

No special labs or equipment needed

The more careful you are, the better it gets

Fast

On order of 1 day to apply

(Standard usability testing may take a week)

Easy to use

Can be taught in 2-4 hours

“DISCOUNT USABILITY” TECHNIQUES

Heuristic Evaluation

Assess interface based on a predetermined list of criteria

Cognitive Walkthroughs

Put yourself in the shoes of a user

Like a code walkthrough

Other, non-inspection techniques are on the rise

e.g., online remote experiments with Mechanical Turk

COGNITIVE WALKTHROUGH

COGNITIVE WALKTHROUGH

Formalized technique for imagining user's thoughts and actions when using an interface:

"Cognitive walkthroughs involve simulating a user's problem-solving process at each step in the human-computer dialog, checking to see if the user's goals and memory for actions can be assumed to lead to the next correct action." (Nielsen, 1992)

COGNITIVE WALKTHROUGH

Given an interface prototype or specification, need:

- A detailed task with a concrete goal,
  ideally motivated by a scenario
- Action sequences for user to complete the task

Ask the following questions for each step:

- Will the users know what to do?
- Will the user notice that the correct action is available?
- Will the user interpret the application feedback correctly?

Record: what would cause problems, and why?

From: Preece, Rogers, Sharp – Interaction Design

COGNITIVE WALKTHROUGH EXAMPLE

Task: Find the call number and location of the latest edition of the book "Interaction Design" by Preece, Rogers & Sharp in the Berkeley library

Typical users: Students who are familiar with the web, but not necessarily with the library or its website

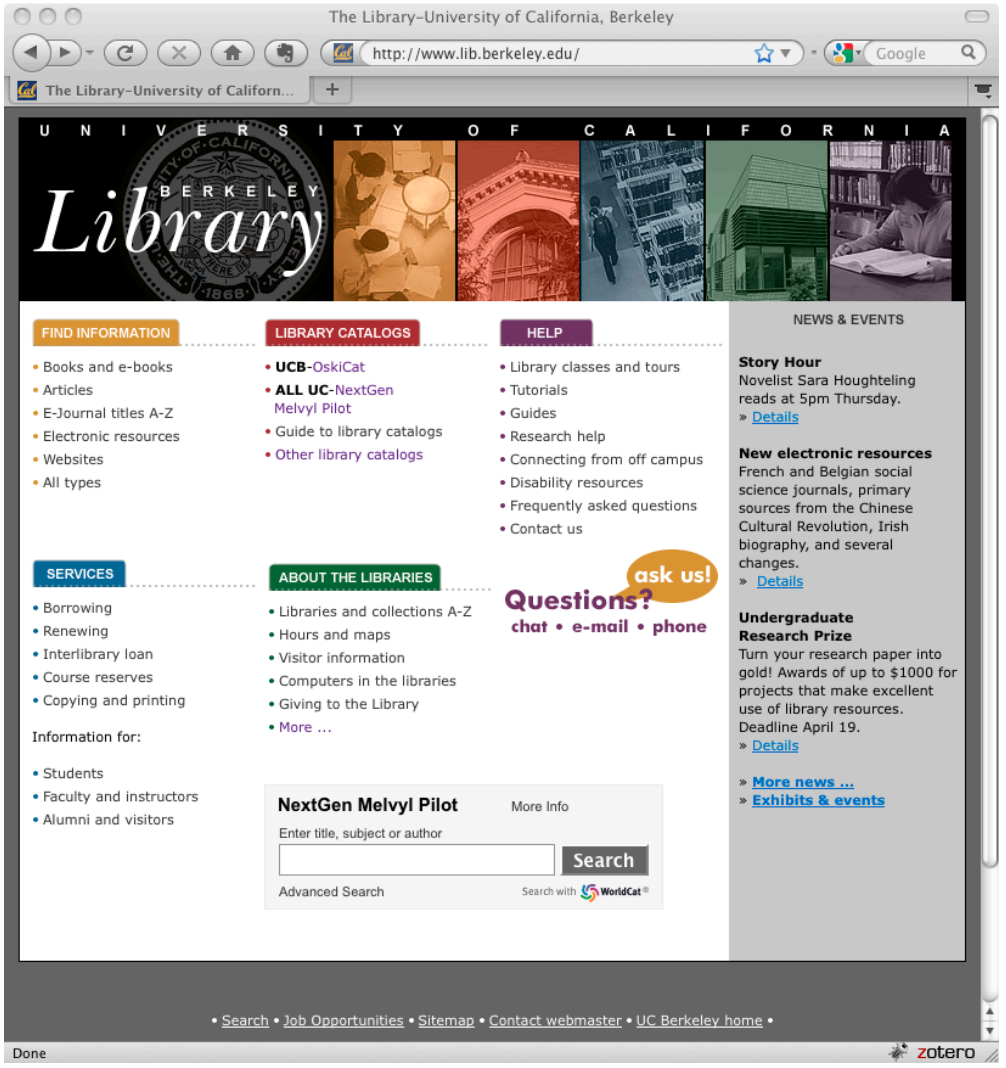
COGNITIVE WALKTHROUGH EXAMPLE

Step1: Select library catalog.

Will the user know what to do?

Will user notice that action is available?

Will user interpret feedback correctly?



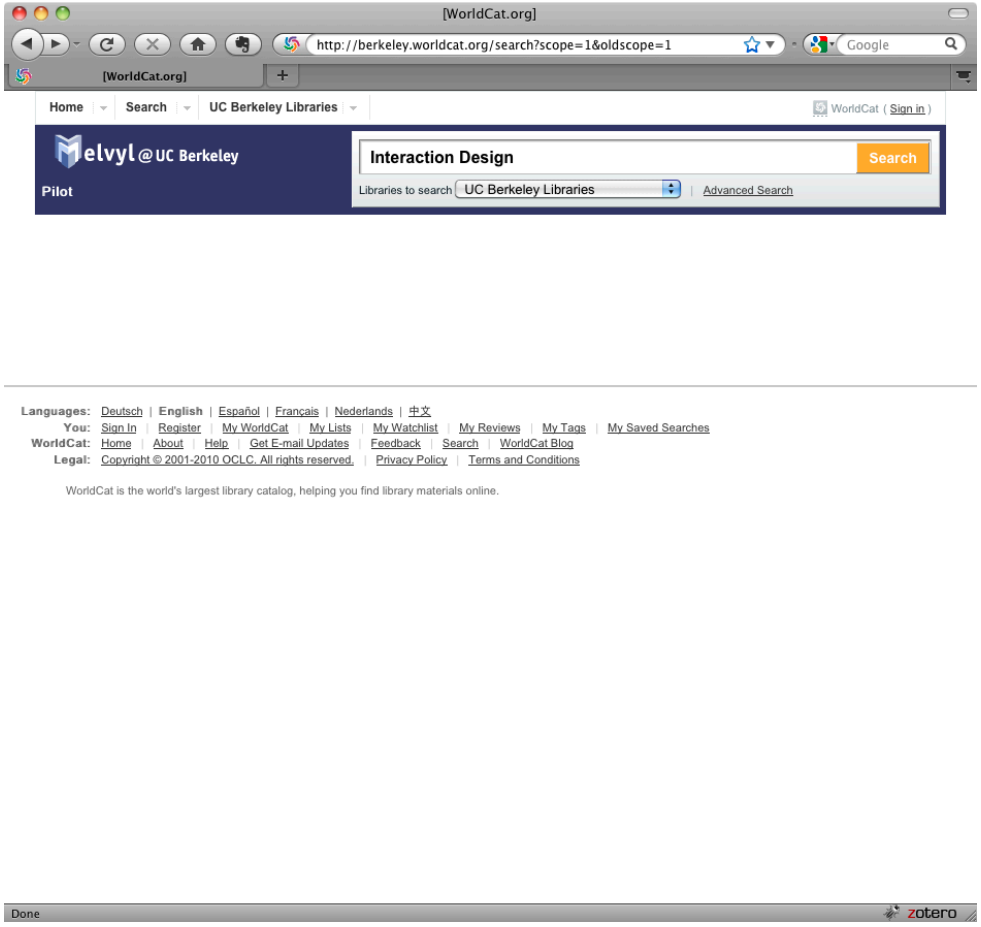
COGNITIVE WALKTHROUGH EXAMPLE

Step 2: Complete the search form

Will the user know what to do?

Will user notice that action is available?

Will user interpret feedback correctly?



COGNITIVE WALKTHROUGH EXAMPLE

Step 3: Locate the right edition, click to detail screen

Will the user know what to do?

Will user notice that action is available?

Will user interpret feedback correctly?

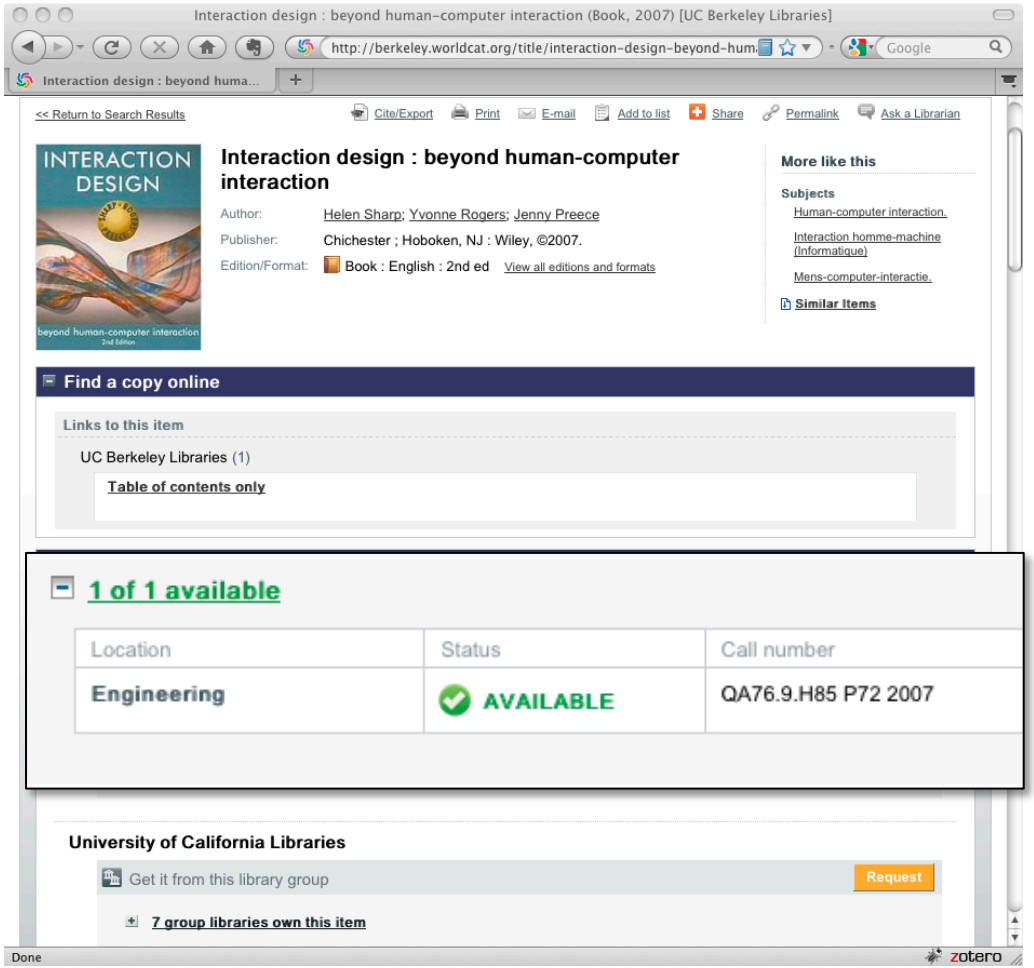
COGNITIVE WALKTHROUGH EXAMPLE

Step 4: Locate call number and library location

Will the user know what to do?

Will user notice that action is available?

Will user interpret feedback correctly?



EMPIRICAL ASSESSMENT: QUALITATIVE

Qualitative: What we've been doing so far

Contextual Inquiry: try to understand user's tasks and conceptual model

Usability Studies: look for critical incidents in interface

Qualitative methods help us:

Understand what is going on

Look for problems

Roughly evaluate usability of interface

EMPIRICAL: QUANTITATIVE STUDIES

Quantitative

Use to reliably measure some aspect of interface

Compare two or more designs on a measurable aspect

Contribute to theory of Human-Computer Interaction

Approaches

Collect and analyze user events that occur in natural use
Controlled experiments

Examples of measures

Time to complete a task, Average number of errors on a task, Users' ratings of an interface*

** You could argue that users' perception of speed, error rates etc is more important than their actual values*

COMPARISON

Qualitative studies

Faster, less expensive → esp. useful in early stages of design cycle

Quantitative studies

Reliable, repeatable result → scientific method

Best studies produce generalizable results

DESIGNING CONTROLLED EXPERIMENTS

STEPS IN DESIGNING AN EXPERIMENT

1. State a lucid, testable hypothesis
2. Identify variables
   (independent, dependent, control, random)
3. Design the experimental protocol
4. Choose user population
5. Apply for human subjects protocol review
6. Run pilot studies
7. Run the experiment
8. Perform statistical analysis
9. Draw conclusions

EXPERIMENT DESIGN

Testable hypothesis

Precise statement of expected outcome

Independent variables (factors)

Attributes we manipulate/vary in each condition

Levels – values for independent variables

Dependent variables (response variables)

Outcome of experiment (measurements)

Usually measure user performance

EXPERIMENT DESIGN

Control variables

Attributes that will be fixed throughout experiment

Confound – attribute that varied and was not accounted for

Problem: Confound rather than independent variables could have caused change in dependent variables

Confounds make it difficult/impossible to draw conclusions

Random variables

Attributes that are randomly sampled

Increases generalizability

VARIABLE TYPES

Nominal: categories with labels, no order

Ordinal: categories with rank order

Continuous:
interval (w/o zero point), ratio (w/ zero point)

COMMON METRICS IN HCI

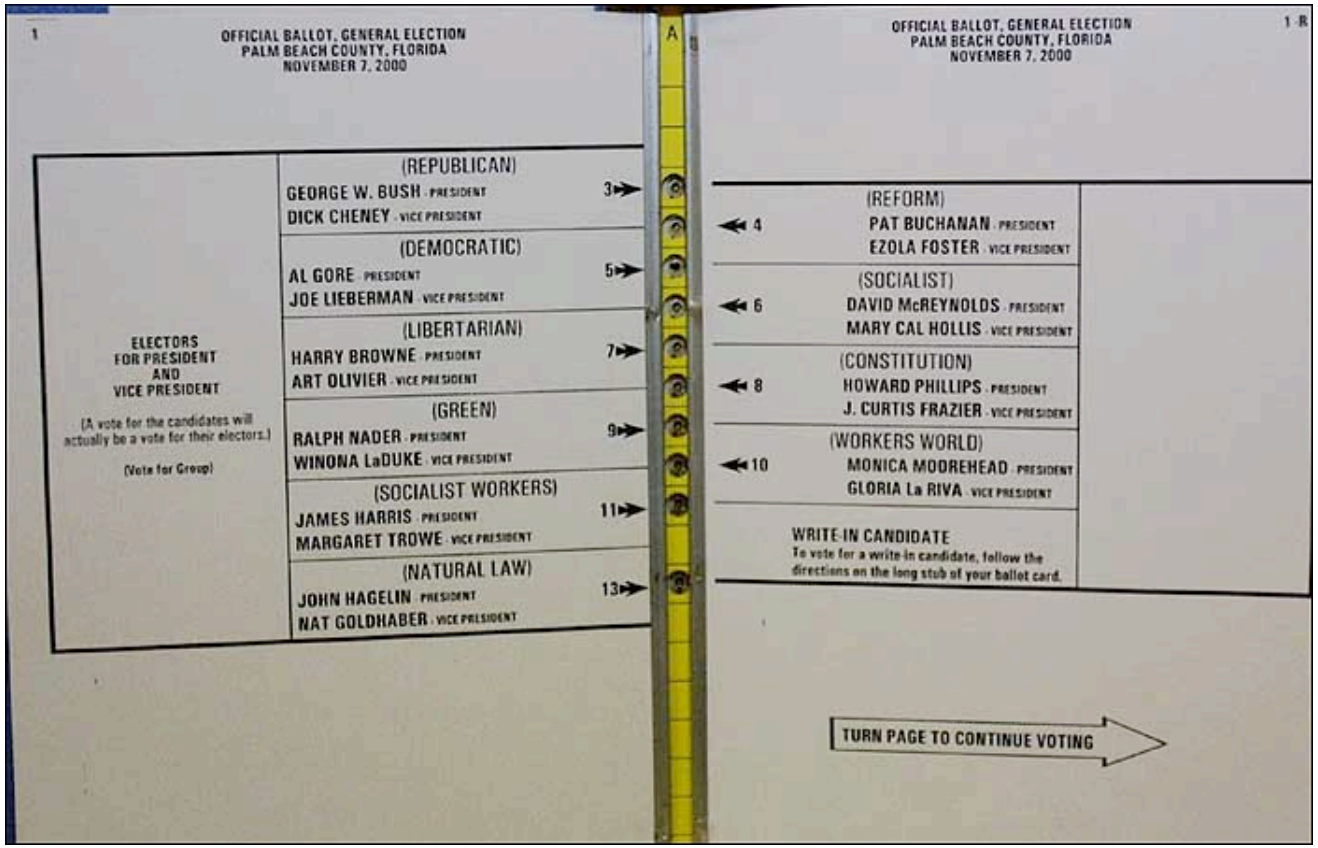
Performance metrics:

- Task success (binary or multi-level)
- Task completion time
- Errors (slips, mistakes) per task
- Efficiency (cognitive & physical effort)
- Learnability

Satisfaction metrics:

- Self-report on ease of use, frustration, etc.

PERFORMANCE METRIC: ERRORS



1

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS FOR PRESIDENT AND VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN) 3
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT

(DEMOCRATIC) 5
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT

(LIBERTARIAN) 7
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT

(GREEN) 9
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT

(SOCIALIST WORKERS) 11
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT

(NATURAL LAW) 13
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT

A

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

1-R

(REFORM) 4
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

(SOCIALIST) 6
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

(CONSTITUTION) 8
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

(WORKERS WORLD) 10
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

TURN PAGE TO CONTINUE VOTING

stcsig.org

media.tbo.com / AP

PERFORMANCE METRIC: LOSTNESS

Smith 1996:

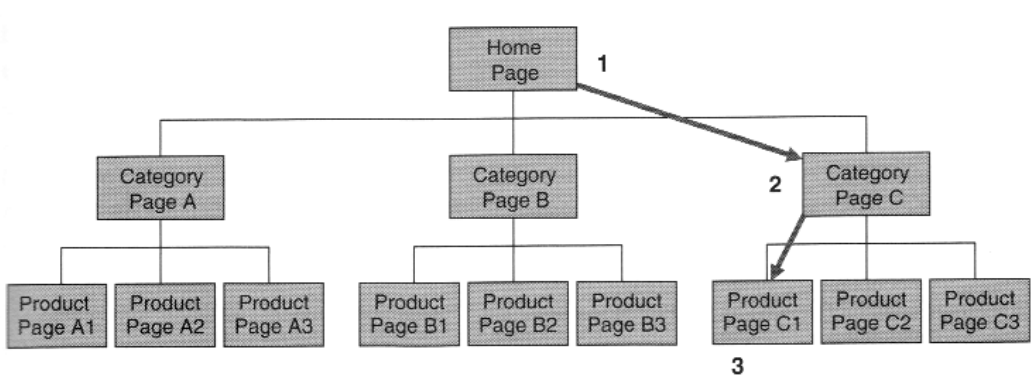
N: # of different pages visited

S: # of total pages visited, incl. revisits

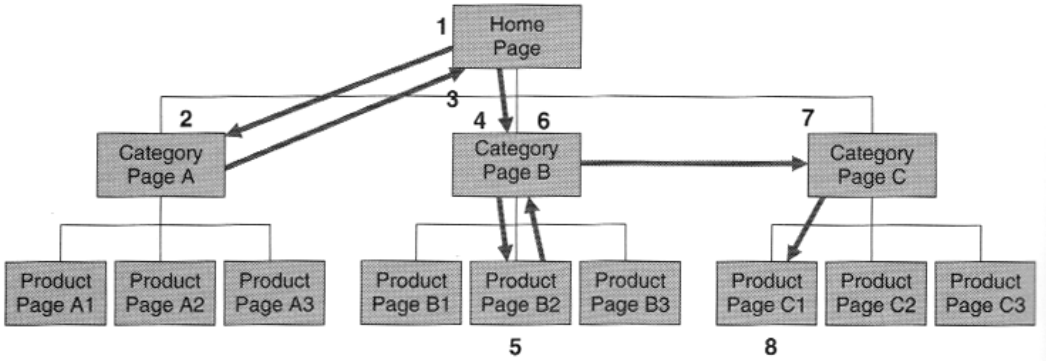
R: minimum # of pages to accomplish task

Lostness = $\text{sqrt}((N/S-1)^2+(R/N-1)^2)$

Smith 1996



Optimum number of steps (three) to accomplish a task that involves finding a target item on Product Page C1 starting from the homepage.



Actual number of steps a participant took in getting to the target item on Product Page C1. Note that each revisit to the same page is counted, giving a total of eight steps.

SATISFACTION METRIC: LIKERT SCALES

Respondents rate their level of agreement to a statement

Likert data is ordinal, not continuous (matters for analysis)!

"Overall, I am satisfied with the ease of completing the tasks in this scenario"

1: Strongly Disagree
2: Disagree
3: Neither agree nor disagree
4: Agree
5: Strongly agree

The Bubble Cursor: Enhancing Target Acquisition by Dynamic Resizing of the Cursor's Activation Area

Tovi Grossman

Ravin Balakrishnan

Dynamic Graphics Project Lab
Department of Computer Science
University of Toronto
www.dgp.toronto.edu

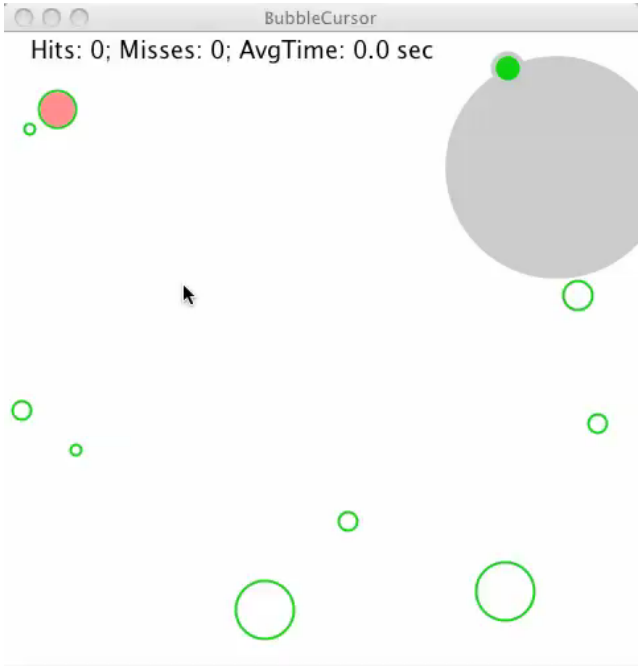
VARIABLES FOR THE BUBBLE CURSOR

Independent variables

Dependent variables

Control variables

Random variables



VARIABLES

Independent variables

Cursor type (bubble, normal, area?)
Target Distance
Target Width

Dependent variables

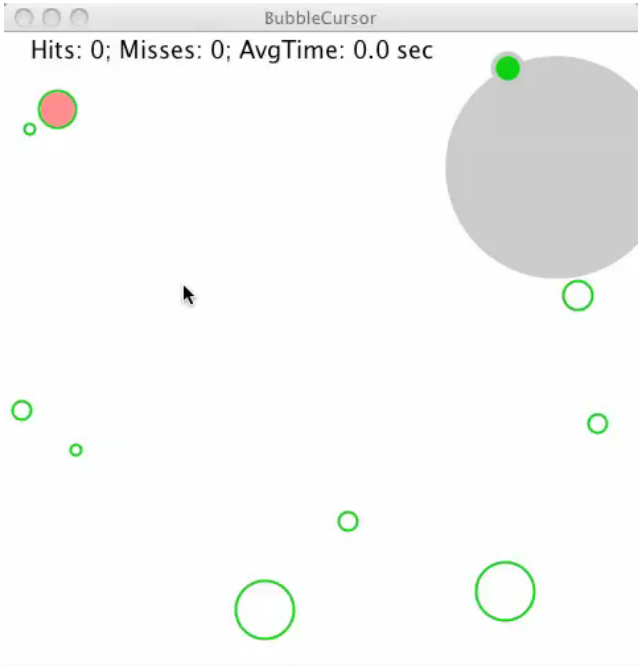
Movement Time
Error Rate
User Satisfaction

Control variables

Color scheme, input device, screen size

Random variables

Location, environment,
Attributes of subjects
Age, gender, handedness, ...



Conducting studies online vs. in person strongly influences which variables are controlled and which are random.

GOALS

Internal validity

Manipulation of IV is cause of change in DV

Requires eliminating confounding variables (turn them into IVs or RVs)

Requires that experiment is replicable

External validity

Results are generalizable to other experimental settings

Ecological validity – results generalizable to real-world settings

Confidence in results

Statistics

EXPERIMENTAL PROTOCOL

What is the task? (must reflect hypothesis!)

What are all the combinations of conditions?

How often to repeat each combination of conditions?

Between subjects or within subjects

Avoid bias (instructions, ordering, ...)

NUMBER OF CONDITIONS

Consider all combinations to isolate effects of each IV (factorial design)

(3 cursor types) * (3 distances) * (3 widths) = 27 combinations

Adding levels or factors can yield lots of combinations!

REDUCING NUMBER OF CONDITIONS

Vary only one independent variable
leaving others fixed

Problem: ?

REDUCING NUMBER OF CONDITIONS

Vary only one independent variable
leaving others fixed

Problem: Will miss effects of interactions

OTHER REDUCTION STRATEGIES

Run a few independent variables at a time

If strong effect, include variable in future studies

Otherwise pick fixed control value for it

Fractional factorial design

Procedures for choosing subset of independent variables to vary in each experiment

CHOOSING SUBJECTS

Pick balanced sample reflecting intended user population

Novices, experts

Age group

Sex

....

Example

12 non-colorblind right-handed adults (male & female)

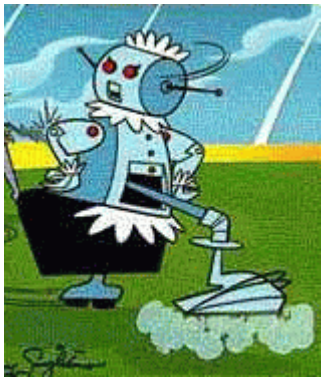
Population group can also be an IV or a controlled variable

What is the disadvantage of making population a controlled var?

BETWEEN SUBJECTS DESIGN

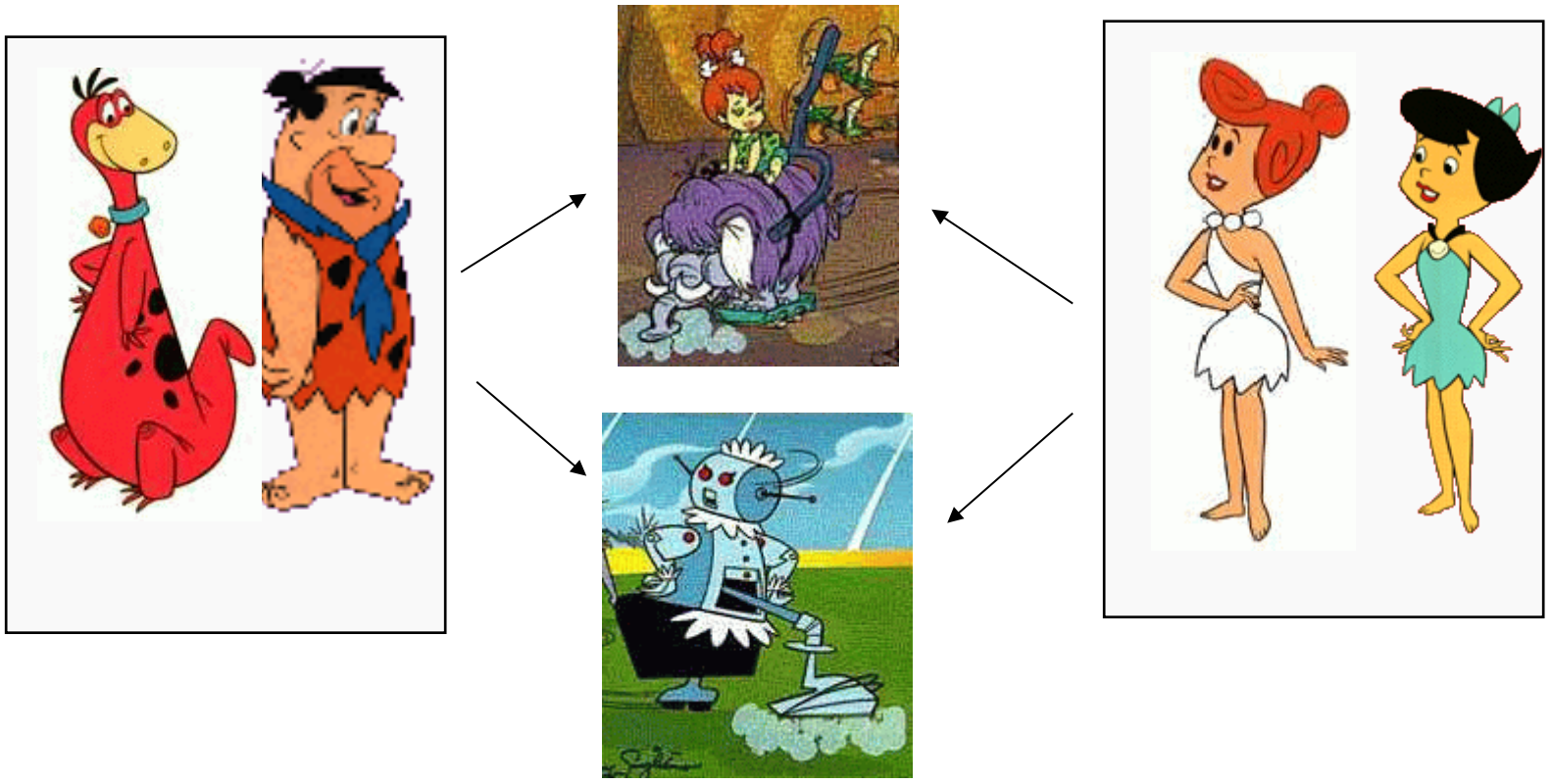
Wilma and Betty use one interface

Dino and Fred use the other

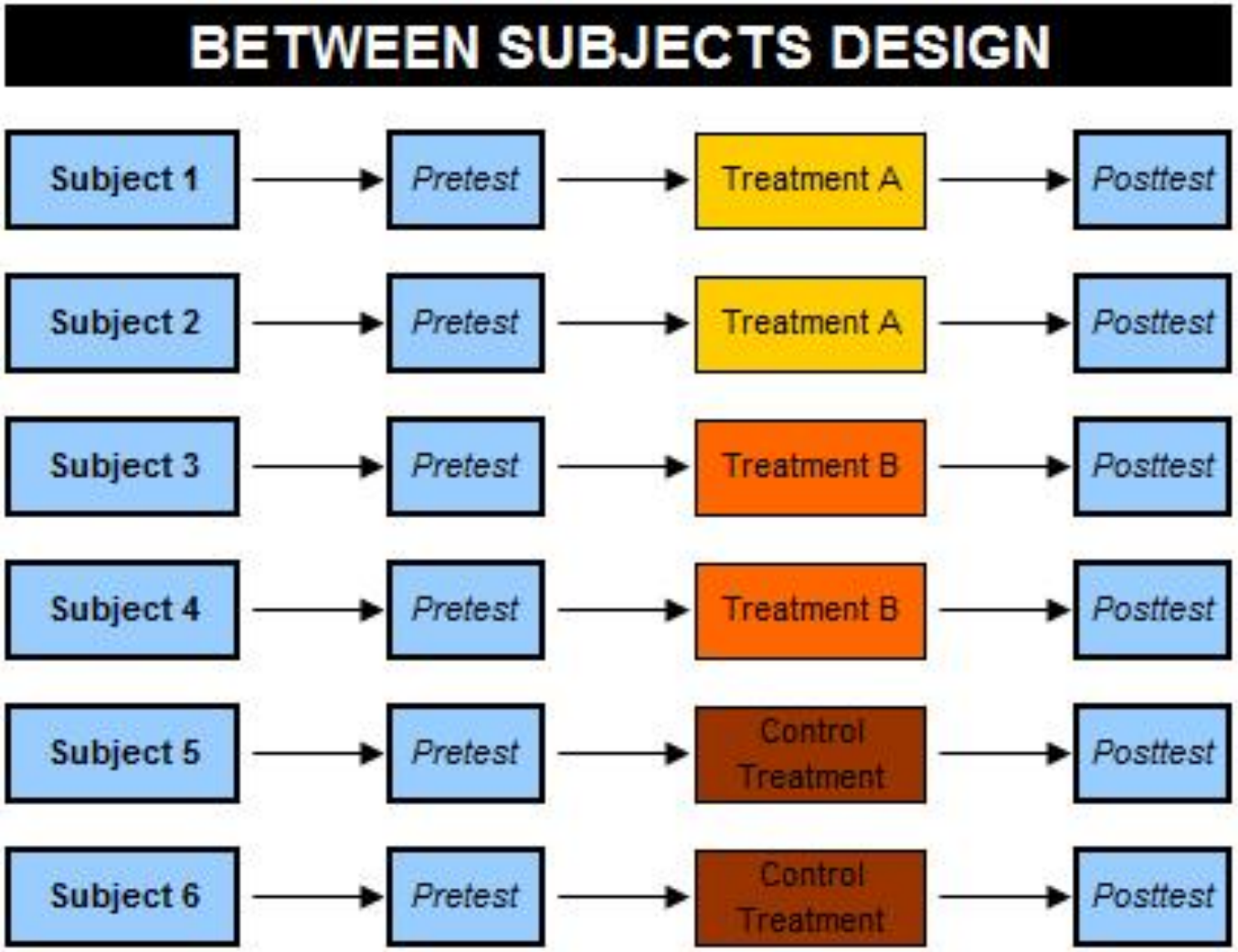


WITHIN SUBJECTS DESIGN

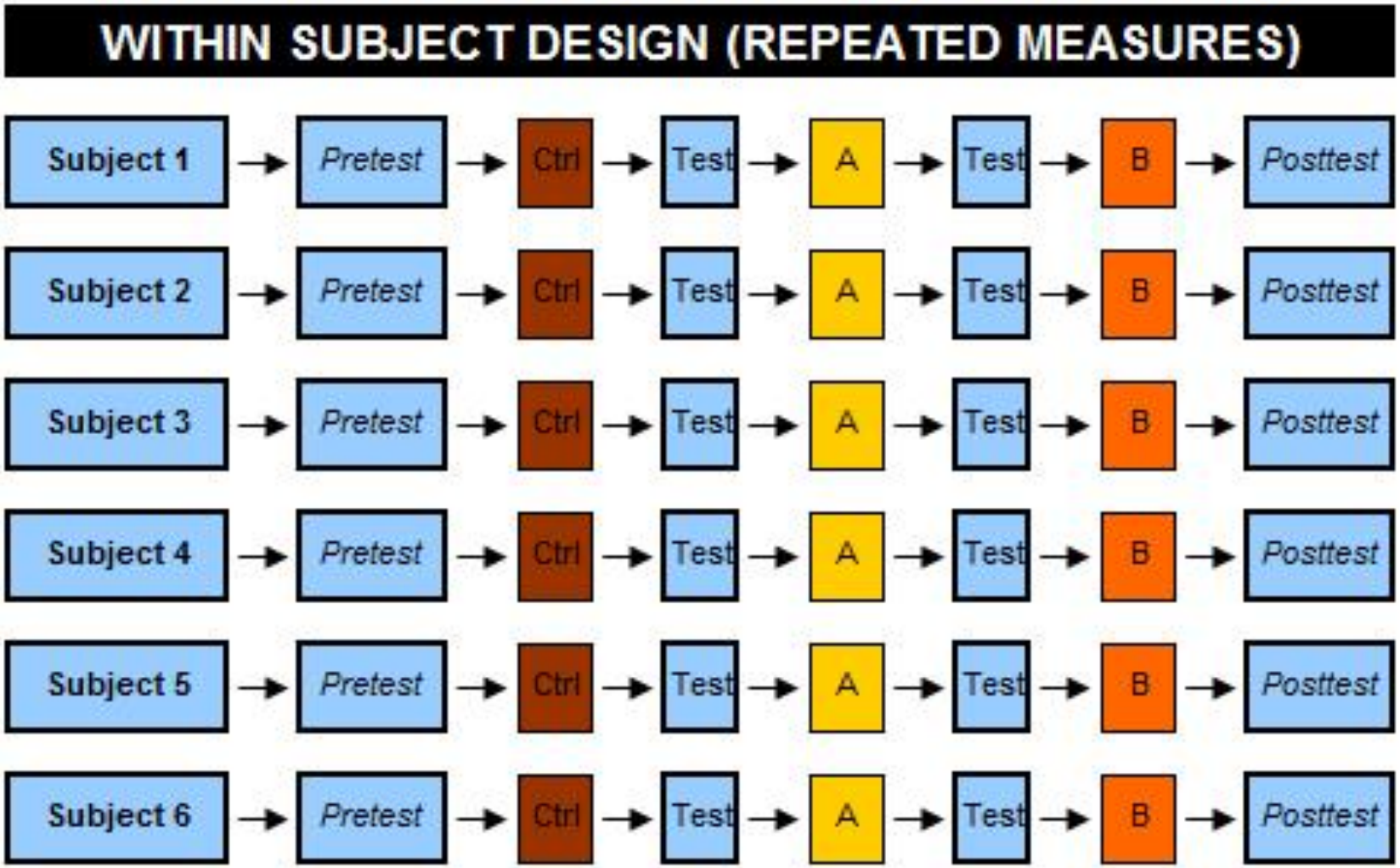
Everyone uses both interfaces



BETWEEN SUBJECTS DESIGN



WITHIN SUBJECTS DESIGN



Within Subjects

A group of people sees the test signs.

Between Subjects

One group of people sees one set of the test signs, and a different group sees another set.

BETWEEN VS. WITHIN SUBJECTS

Between subjects

Each participant uses one condition

+/- Participants cannot compare conditions

+ Can collect more data for a given condition

- Need more participants

Within subjects

All participants try all conditions

+ Compare one person across conditions to isolate effects of individual diffs

+ Requires fewer participants

- Fatigue effects

- Bias due to ordering/learning effects

WITHIN SUBJECTS: ORDERING EFFECTS

In within-subjects designs ordering of conditions is a variable that can confound results

Why?

Turn it into a random variable

Randomize order of conditions across subjects

Counterbalancing (ensure all orderings are covered)

Latin square (partial counterbalancing)

…

RUN THE EXPERIMENT

Always pilot it first!

Reveals unexpected problems

Can't change experiment design after starting it

Always follow same steps – use a checklist

Get consent from subjects

Debrief subjects afterwards