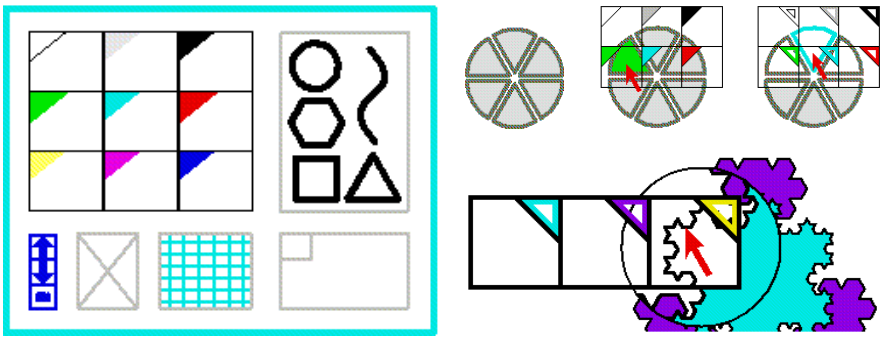
Quantitative Evaluation

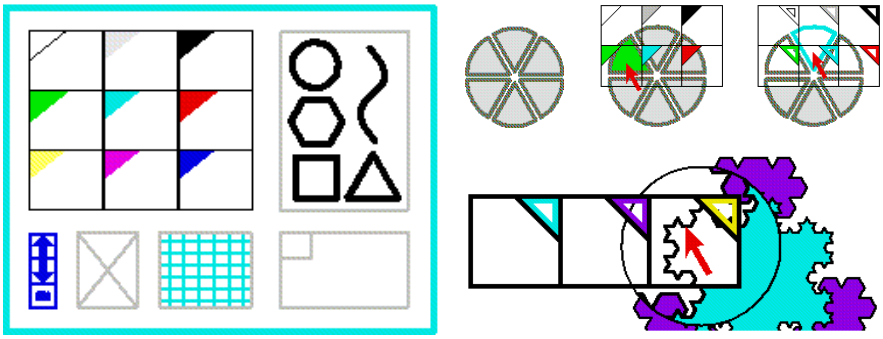
CS160: User Interfaces
Maneesh Agrawala

Slides based on those of Francois Guimbretiere, John Canny and Marti Hearst



Toolglass and Magic Lenses: The See-Through Interface [Bier et al. 93]

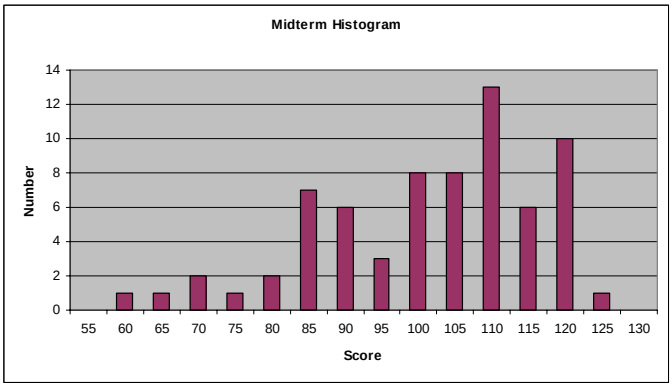
VIDEO



Toolglass and Magic Lenses: The See-Through Interface [Bier et al. 93]

Midterm

Handed back last week



Mean: 99.42
Stdev: 15.34

Upcoming Schedule

Team Assessment (due before class today)

Pilot User Study (due Monday Nov 13 before class)

- 3 users will test 3 tasks (one easy, one medium, one hard)
- Finish necessary implementation
  - WOZ is fine – you will probably need to build interface to ease the job of the person acting as computer
  - Canned functionality is **not** ok
- Testing takes time so start early

Review: Discount Usability Eng.

- Walkthroughs
  - Put yourself in the shoes of a user
  - Like a code walkthrough
- Action analysis
  - GOMS (add times to formal action analysis)
- ***Heuristic evaluation***

- Low-fi testing
- On-line, remote usability tests

Review: Heuristic Evaluation

"Rules of thumb" describing features of usable systems
- Can be used as design principles
- Can be used to evaluate a design

Example: *Minimize users' memory load*

Pros and cons
- Easy and inexpensive
  - Performed by experts
  - No users required
  - Catch many design flaws
- More difficult than it seems
  - Not a simple checklist
  - Cannot assess how well the interface will address user goals

Topics

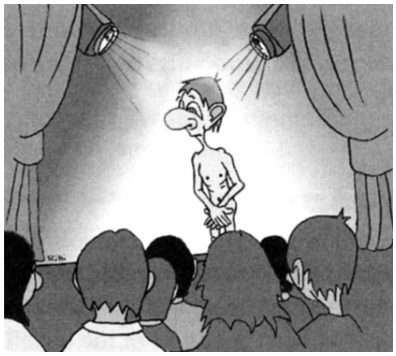
- Managing study participants (qual. and quant. studies)
- Why do we conduct quantitative studies?
- Designing controlled experiments

Managing Study Participants

The Participants' Standpoint

Testing is a distressing experience

- Pressure to perform
- Feeling of inadequacy
- Looking like a fool in front of your peers, your boss,…



(from "Paper Prototyping" by Snyder)

Treating Subjects With Respect

Follow human subject protocols

- Individual test results will be kept confidential
- Users can stop the test at any time
- Users are aware (and understand) the monitoring technique
- Their performance will have not implication on their life
- Records will be made anonymous
  - Videos

Use standard informed consent form

- Especially for quantitative tests
- Be aware of legal requirements

Conducting the Experiment

Before the experiment

- Have them read and sign the consent form
- Explain the goal of the experiment
  - In a way accessible to users
  - Be careful about the demand characteristic
    - Participants biased towards experimenter's hypothesis
- Answer questions

During the experiment

- Stay neutral
- Never indicate displeasure with users performance

After the experiment

- Debrief users
  - Inform users about the goal of the experiment
- Answer any questions they have

Managing Subjects

Don't waste users time

- Use pilot tests to debug experiments, questionnaires, etc…
- Have everything ready before users show up

Make users comfortable

- Keep a relaxed atmosphere
- Allow for breaks
- Pace tasks correctly
- Stop the test if it becomes too unpleasant

Ethics: Stanford Prison Experiment

1971 Experiment by Phil Zimbardo at Stanford

- 24 Participants – half prisoners, half guards ($15 a day)
- Basement of Stanford Psychology bldg turned into mock prison
- Guards given batons, military style uniform, mirror glasses,…
- Prisoners wore smocks (no underwear), thong sandals, pantyhose caps

Experiment quickly got out of hand

- Prisoners suffered and accepted sadistic treatment
- Prison became unsanitary/inhospitable
- Prisoner riot put down with use of fire extinguishers
- Guards volunteered to work extra hours

Zimbardo terminated experiment early

- Grad student Christina Maslach objected to experiment
- Important to check protocol with ethics review boards

[from Wikipedia]

Ethics

Was it useful?

- "…that's the most valuable kind of information that you can have - and that certainly a society needs it" (Zimbardo)

Was it ethical?

- Could we have gather this knowledge by other means?



http://www.prisonexp.org/slide-42.htm

Why Quantitative Studies?

Qualitative Studies

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

Quantitative Studies

Quantitative

- Use to reliably measure some aspect of interface
- Compare two or more designs on a measurable aspect

Approaches

- Collect and analyze user events that occur in natural use
  - mouse clicks, key presses
- Controlled experiments

Examples of measures

- Time to complete a task
- Average number of errors on a task
- Users' ratings of an interface *
  - Ease of use, elegance, performance, robustness, speed,…

* 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
- In real-world design quant. study not always necessary

Quantitative studies

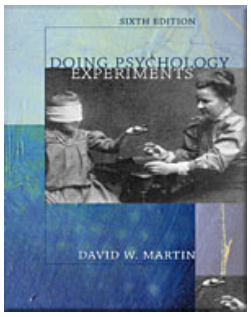
- Reliable, repeatable result → scientific method
- Best studies produce generalizable results

Pilot User Study Assignment

You will conduct a **qualitative** study

- We don't have time or subjects for quantitative studies
- But you should do a little quantitative analysis
  - What are your measures?
  - Compute summary statistics (mean, stdev)
  - Do you have independent, dependent, and control variables?

Designing Controlled Experiments

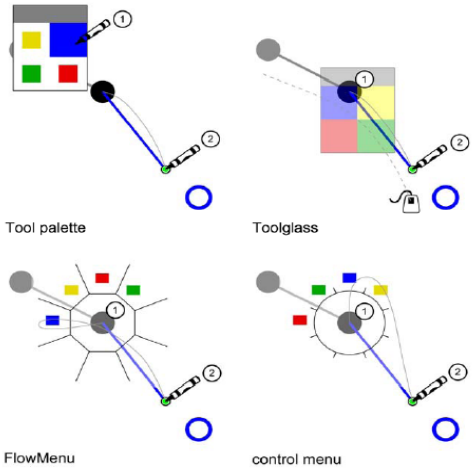


Doing Psychology Experiments
David W. Martin

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

Example: Menu Selection

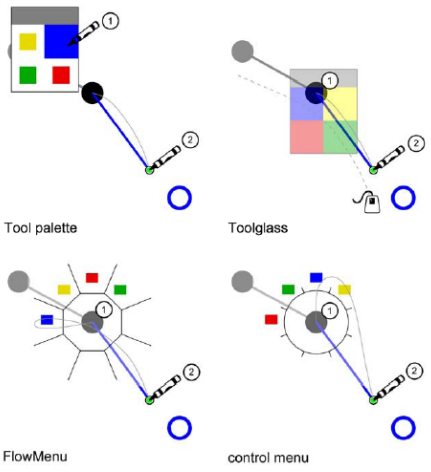


[Guimbretiere et al. 03]

Lucid, Testable Hypothesis

Because users must reach for it, tool palette will be slower

Other hypotheses?



Experiment Design

Testable hypothesis
- Precise statement of expected outcome

Factors (independent variables)
- Attributes we manipulate/vary in each condition
- Levels – values for independent variables

Response variables (dependent variables)
- Outcome of experiment (measurements)
- Usually measure user performance
  - Time
  - Errors

Experiment Design

Control variables
- Attributes that will be fixed throughout experiment
- Confound – attribute that varied and was not accounted for
  - Problem: Confound rather than IV could have caused change in DVs
- Confounds make it difficult/impossible to draw conclusions

Random variables
- Attributes that are randomly sampled
- Increases generalizability

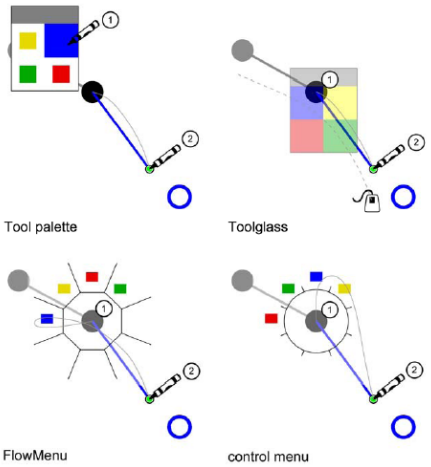
Variables

Independent variables

Dependent variables

Control variables

Random variables



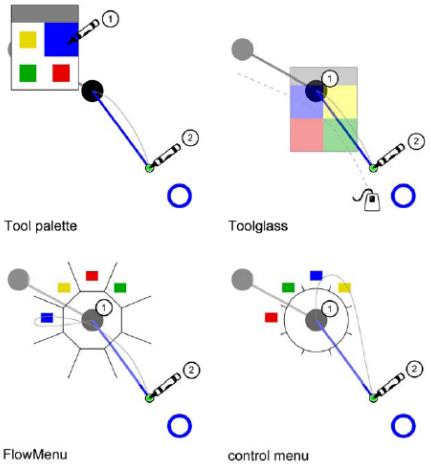
Variables

Independent variables
- Menu type (4 choices)
- Device type (2 choices) ?

Dependent variables
- Time
- Error rate
- User satisfaction

Control variables
- Location/environment …
- Device type ?

Random variables
- Attributes of subjects
  - Age, sex, …



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

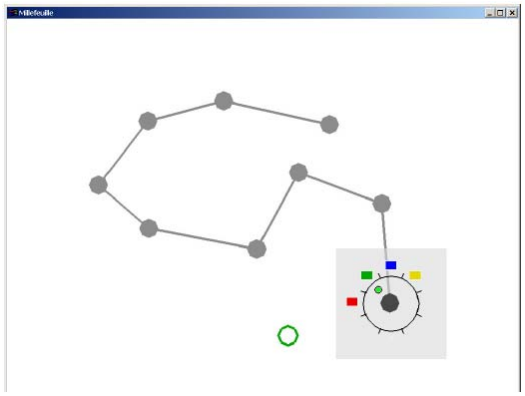
External validity
- Results are generalizable to other experimental settings
- ***Ecological validity*** – results generalizable to real-world settings

Confidence in results
- Statistics

Experimental Protocol

- What is the task?
- What are all the combinations of conditions?
- How often to repeat each combination of conditions?
- Between subjects or within subjects
- Avoid bias (instructions, ordering, …)

Task: Must Reflect Hypothesis



- Connect the dots choosing the given color for each one.
- Connected dots filled in gray. Next dot is open in green.

Number of Conditions

Consider all combinations to isolate effects of each IV (factorial design)
(4 Menu types) * (2 Device types) = 8 combinations

- Tool Palette Pen
- Tool Palette Mouse
- Tool Glass Pen
- Tool Glass Mouse
- Flow Menu Pen
- Flow Menu Mouse
- Control Menu Pen
- Control Menu Mouse

Adding levels or factors can yield lots of combinations!

Reducing Num. of Conditions

Vary only one independent variable leaving others fixed

Problem: ?

Reducing Num. 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)

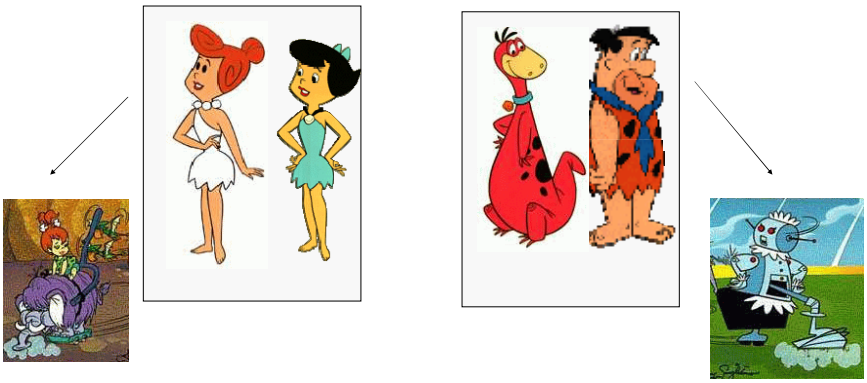
Population group can also be an IV or a controlled variable

- What is the disadvantage of making population a controlled var?
- What are the pros/cons of making population an IV?

Between Subjects Design

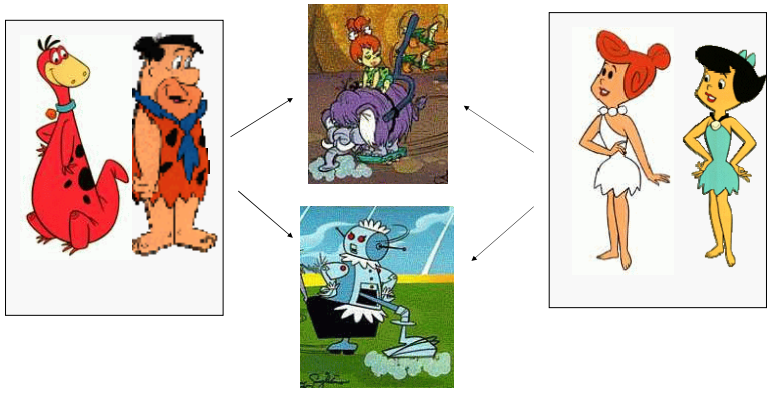
Wilma and Betty use one interface

Dino and Fred use the other



Within Subjects Design

Everyone uses both interfaces



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)
- …

Menu selection example: Within-subjects, each subject tries each condition multiple times, ordering counterbalanced

Run the Experiment

Always pilot it first!
- Reveals unexpected problems
- Can't change experiment design after starting it

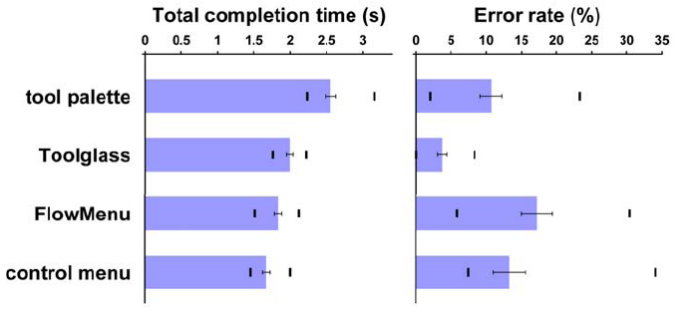
Always follow same steps – use a checklist

Get consent from subjects

Debrief subjects afterwards

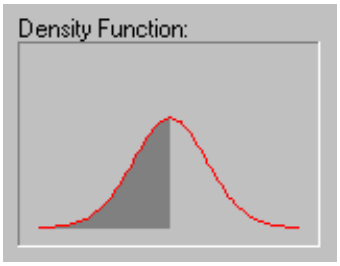
Results: Statistical Analysis

Compute central tendencies (descriptive summary statistics) for each independent variable
- Mean
- Standard deviation



Normal Distributions

Often DVs are assumed to have a Normal distribution



Completely characterized by ***mean*** and ***variance*** (mean squared deviation from the mean).

Are the Results Meaningful?

Hypothesis testing

- **Hypothesis:** Manipulation of IV effects DV in some way
- **Null hypothesis:** Manipulation of IV has no effect on DV
- Null hypothesis assumed true unless statistics allow us to reject it

Statistical significance (p value)

- Likelihood that results are due to chance variation
- $p < 0.05$ usually considered significant (Sometimes $p < 0.01$)
  - Means that < 5% chance that null hypothesis is true

Statistical tests

- T-test (1 factor, 2 levels)
- Correlation
- ANOVA (1 factor, > 2 levels, multiple factors)
- MANOVA (> 1 dependent variable)

Explaining Psychological Statistics
Barry H. Cohen

T-test

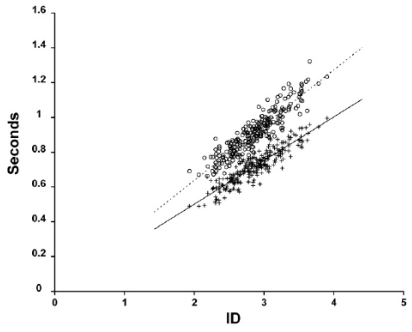
Compare means of 2 groups
- Null hypothesis: No difference between means

Assumptions
- Samples are normally distributed
  - Very robust in practice
- Population variances are equal (between subjects tests)
  - Reasonably robust for differing variances
- Individual observations in samples are independent
  - Extremely important!

Correlation

Measure extent to which two variables are related
- Does not imply cause and effect
  - Example: Ice cream eating and drowning
- Need a large enough sample size

Regression
- Compute the "best fit"
  - linear
  - logistic
  - …



ANOVA

Single factor analysis of variance (ANOVA)
- Compare manes for 3 or more levels of a single independent variable

Multi-Way Analysis of variance (n-Way ANOVA)
- Compare more than one independent variable
- Can find interactions between independent variables

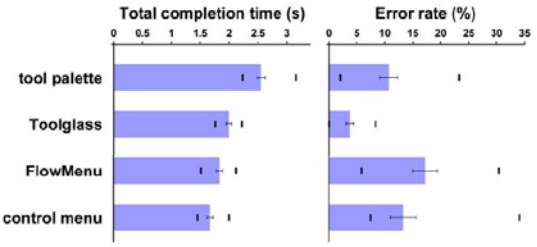
Repeated measures analysis of variance (RM-ANOVA)
- Use when > 1 observation per subject (within subjects expt.)

Multi-variate analysis of variance (MANOVA)
- Compare between more than one dependent var.

ANOVA tests whether means differ, but does not tell us which means differ – for this we must perform pairwise t-tests

Which should we use for the menu selection example?

Menu Selection Example



RM-ANOVA → means for completion times were significantly different ($F(3,33) = 73.4$, $p < .0005$)

- Tool palette significantly slower than others ($p < .0001$ in all cases)
- Control menu faster than FlowMenu but not sig ($p = .2$)
- FlowMenu faster than Toolglass ($p < .01$)
- Control menu faster than Toolglass ($p < .0005$)

Separate analysis for error rates