Quantitative Evaluation and Help

CS160: User Interfaces

John Canny

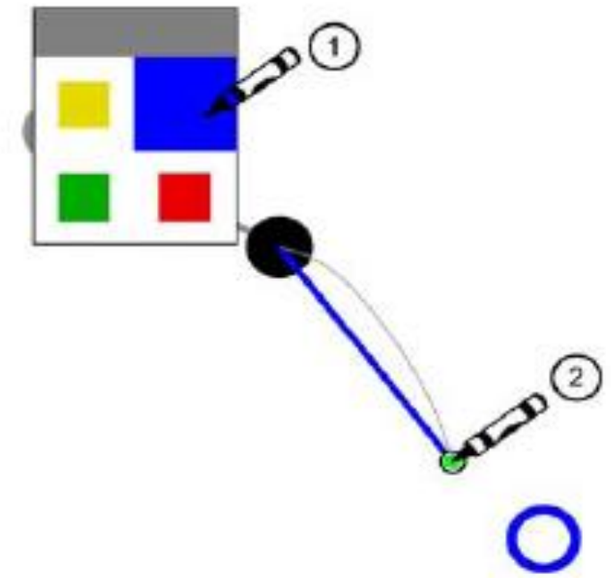
Review

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

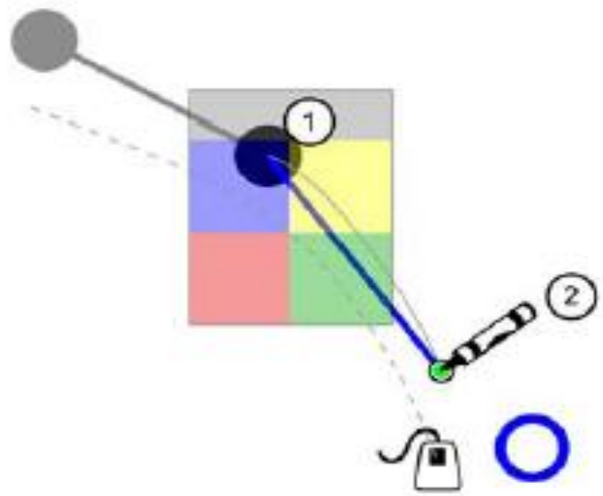
Topics

- Statistical analysis
- Help systems

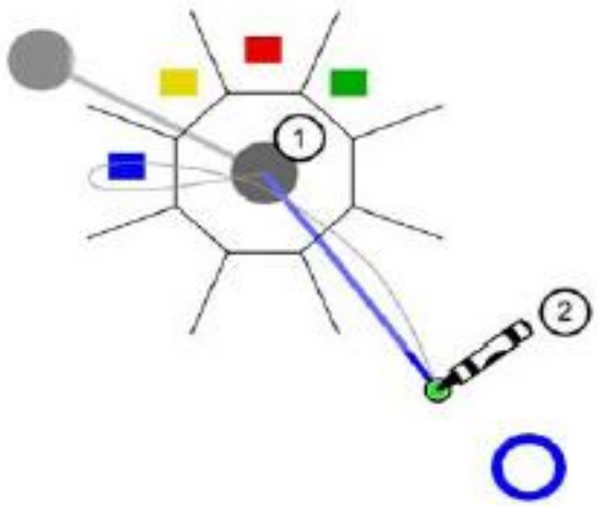
Example: Menu Selection



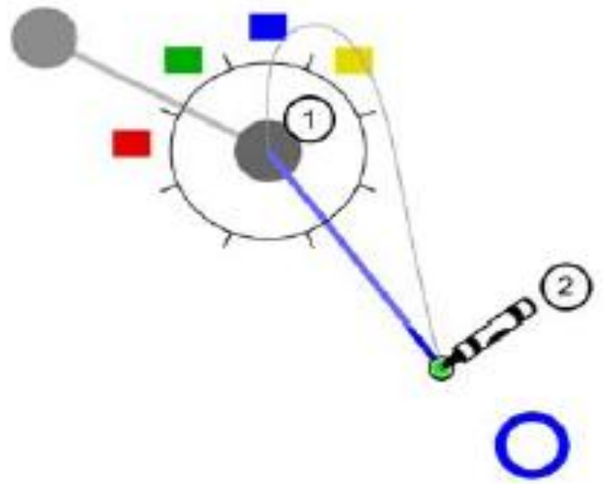
Tool palette



Toolglass



FlowMenu



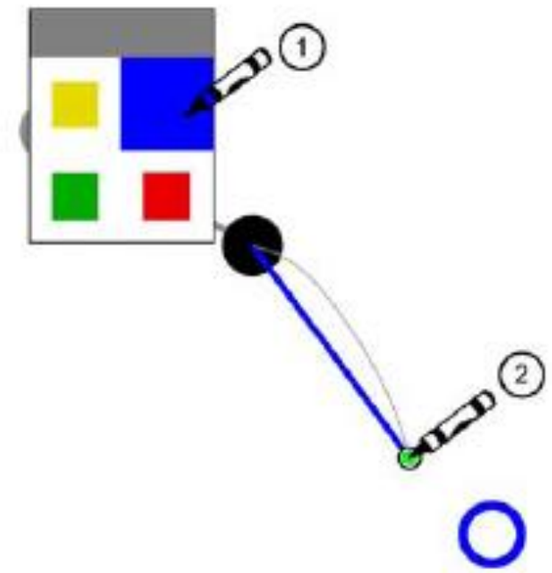
control menu

[Guimbretiere et al. 03]

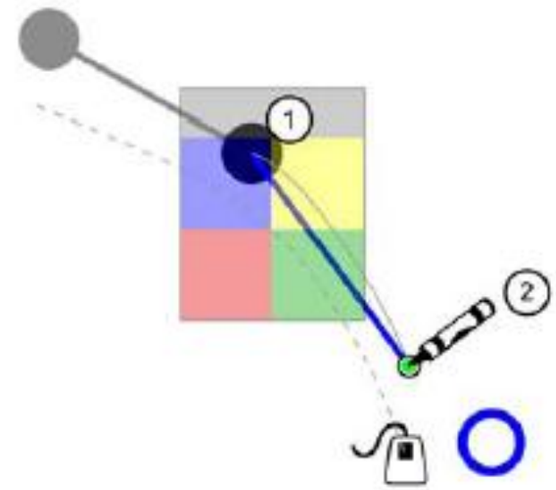
Testable Hypothesis

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

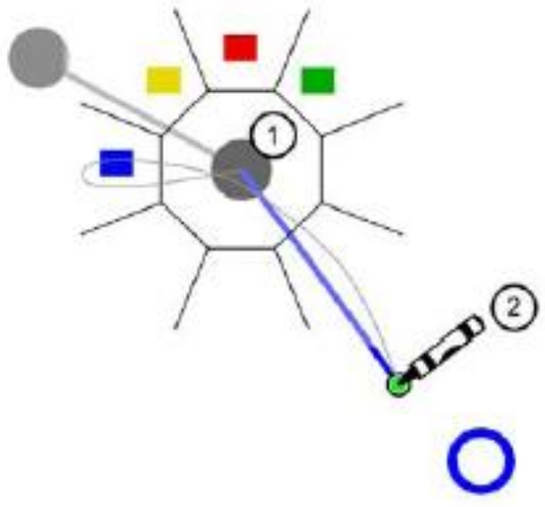
Other hypotheses?



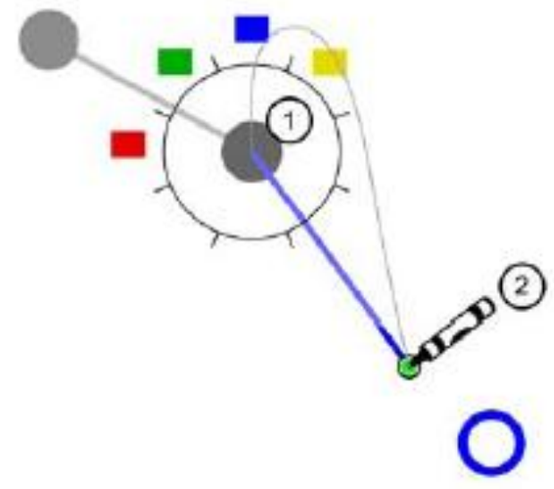
Tool palette



Toolglass



FlowMenu



control menu

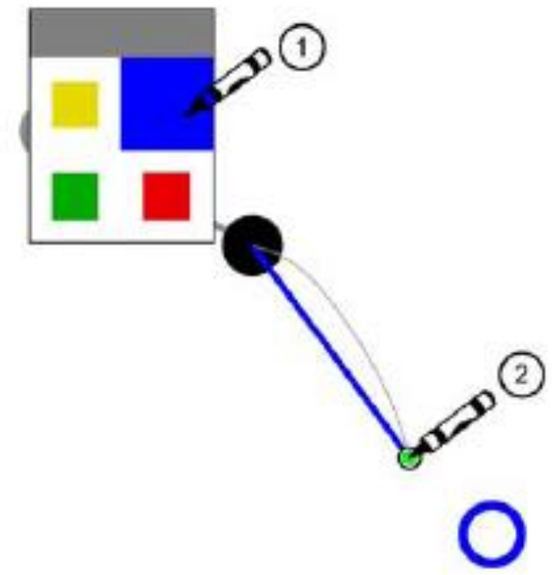
Variables

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

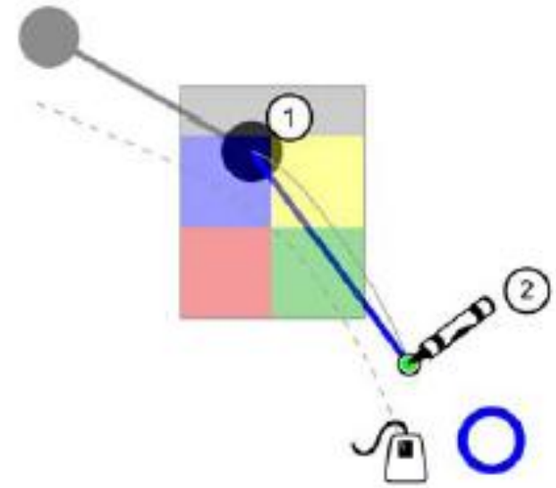
Dependent variables
- Time
- Error rate
- User satisfaction

Control variables
- Location/environment …
- Device type ?

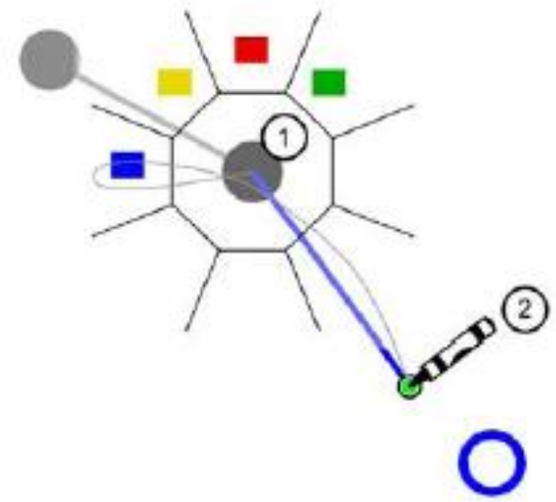
Random variables
- Attributes of subjects
  - Age, sex, …



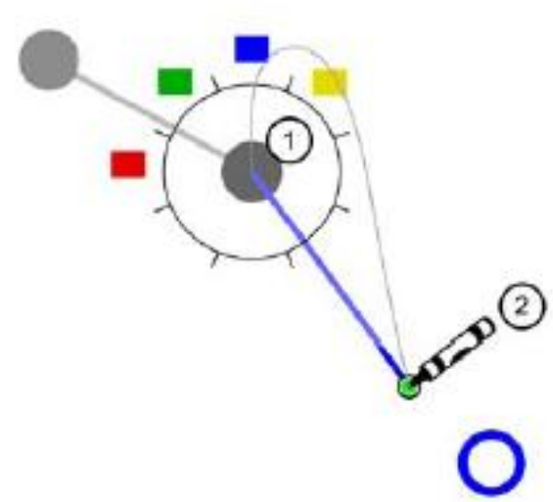
Tool palette



Toolglass



FlowMenu



control menu

Number of Conditions

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

(4 Menu types) * (2 Device types) = 8 combinations

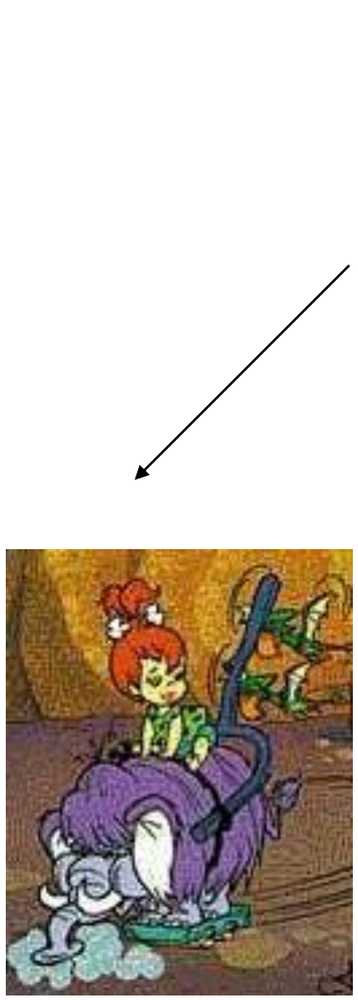
- Tool Palette Pen
- Tool Palette Mouse
- Tool Glass Pen
- Tool Glass Mouse
- Flow Menu Pen
- Flow Menu Mouse
- Control Menu Pen
- Control Menu Mouse

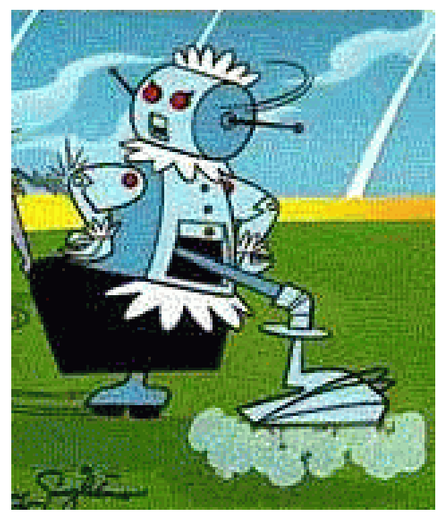
Adding levels or factors can yield lots of combinations!

Between Subjects Design

Wilma and Betty use one interface

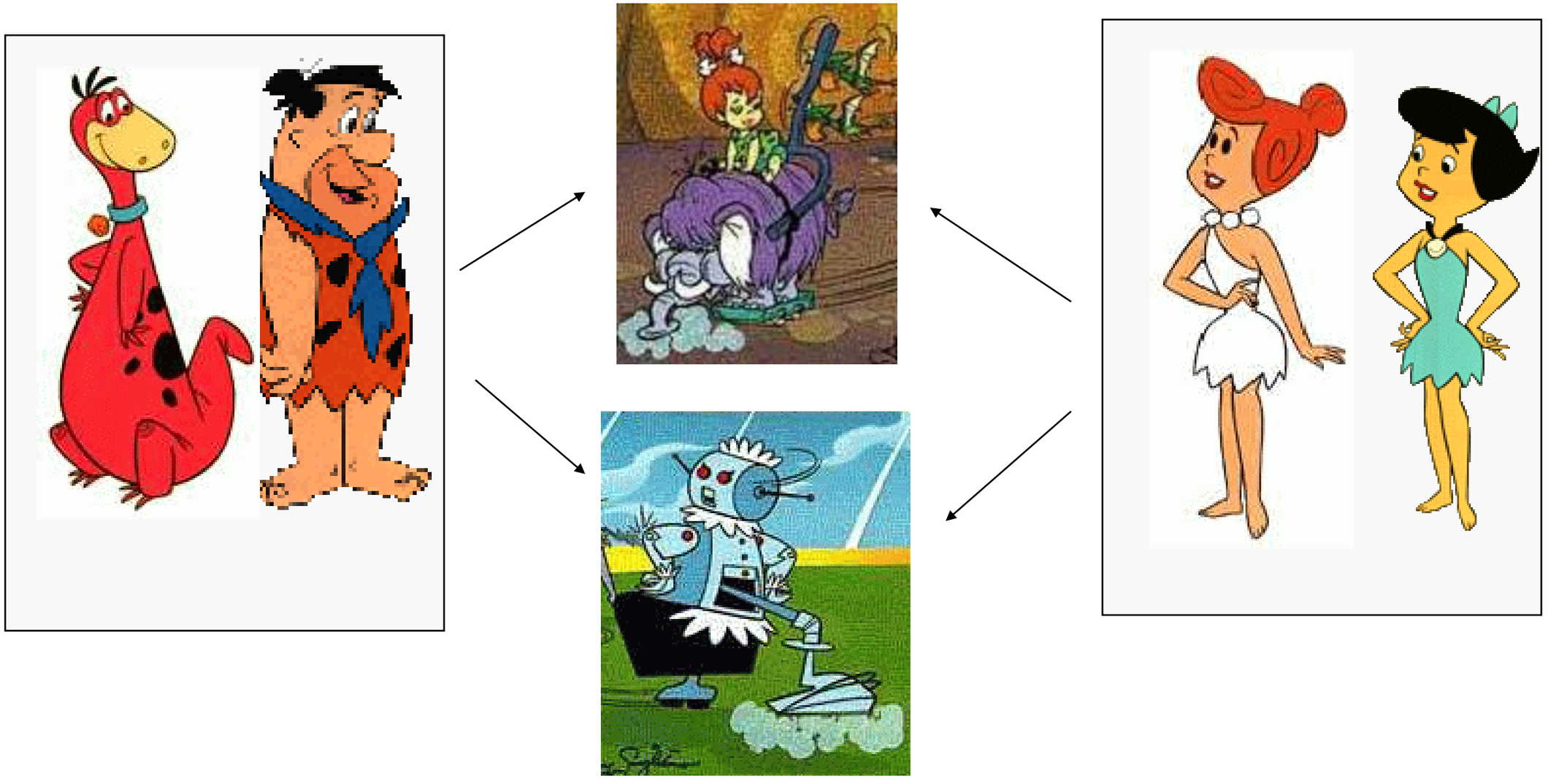
Dino and Fred use the other





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)
- …

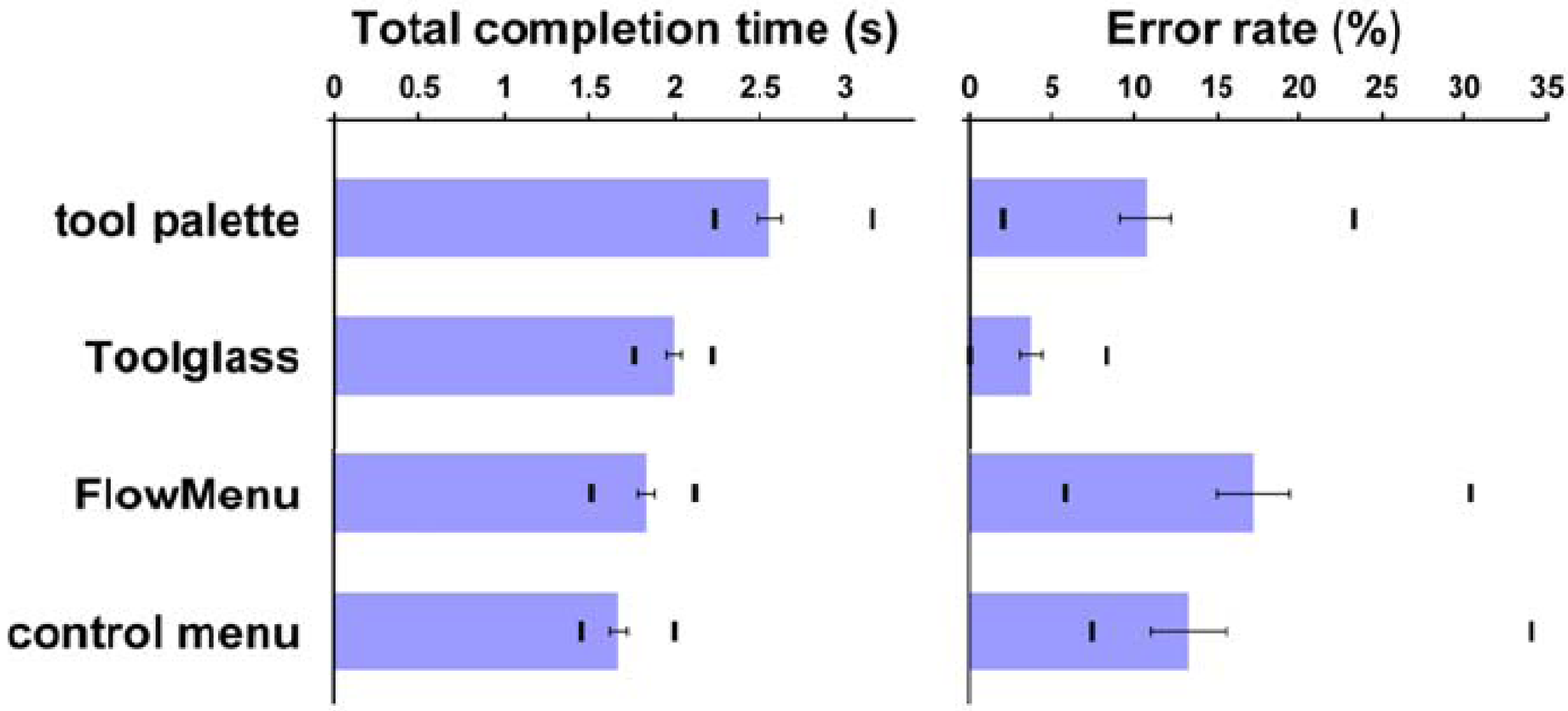
If there are 8 conditions, how many orders would we need to exactly counterbalance?

Solution: choose random orders each time.

Results: Statistical Analysis

Compute central tendencies (descriptive summary statistics) for each independent variable

- Mean
- Standard deviation



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 < 5% chance of the test succeeding given that null hypothesis is true

Statistical tests

- T-test
- Correlation
- ANOVA
- MANOVA

Explaining Psychological Statistics
Barry H. Cohen

T-test

Compare means of 2 groups

- Null hypothesis: No difference between means

Assumptions

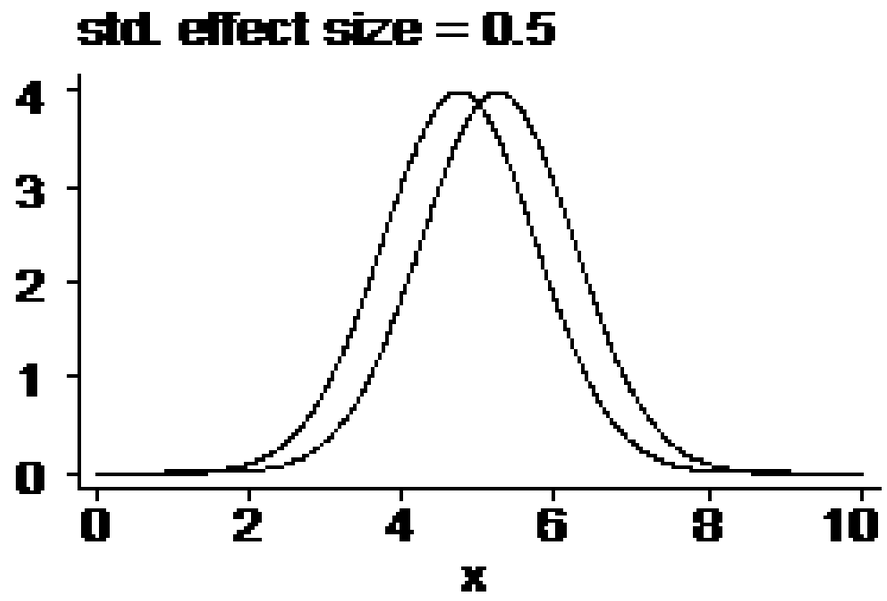
- Samples are normally distributed
  - Very robust in practice
- Population variances are equal (between subjects tests)
  - Reasonably robust for differing variances
  - Try taking logs for variances that scale with mean
- Individual observations in samples are independent
  - Extremely important!

Between Subjects T-test

Let X be the scores of subjects in group A, and Y be the scores of subjects in group B, we use a two-sided t-test. To invoke it in Matlab, do

ttest2(X, Y) (assumes Statistics toolbox)

Where s is the significance, will return the result of the test (pass/fail), the p-value and the value of the t-statistic.



Within Subjects T-test

There is only one group of subjects but if X is their scores on test A and Y is their scores on test B, then X and Y are not independent. Therefore we cannot use the previous two-sided t-test.

Instead we can define the difference Z = X – Y whose components are independent, and ask if Z is significantly different from 0.

ttest(Z) (Matlab)

Returns the result of the test, p-value and t-statistic.

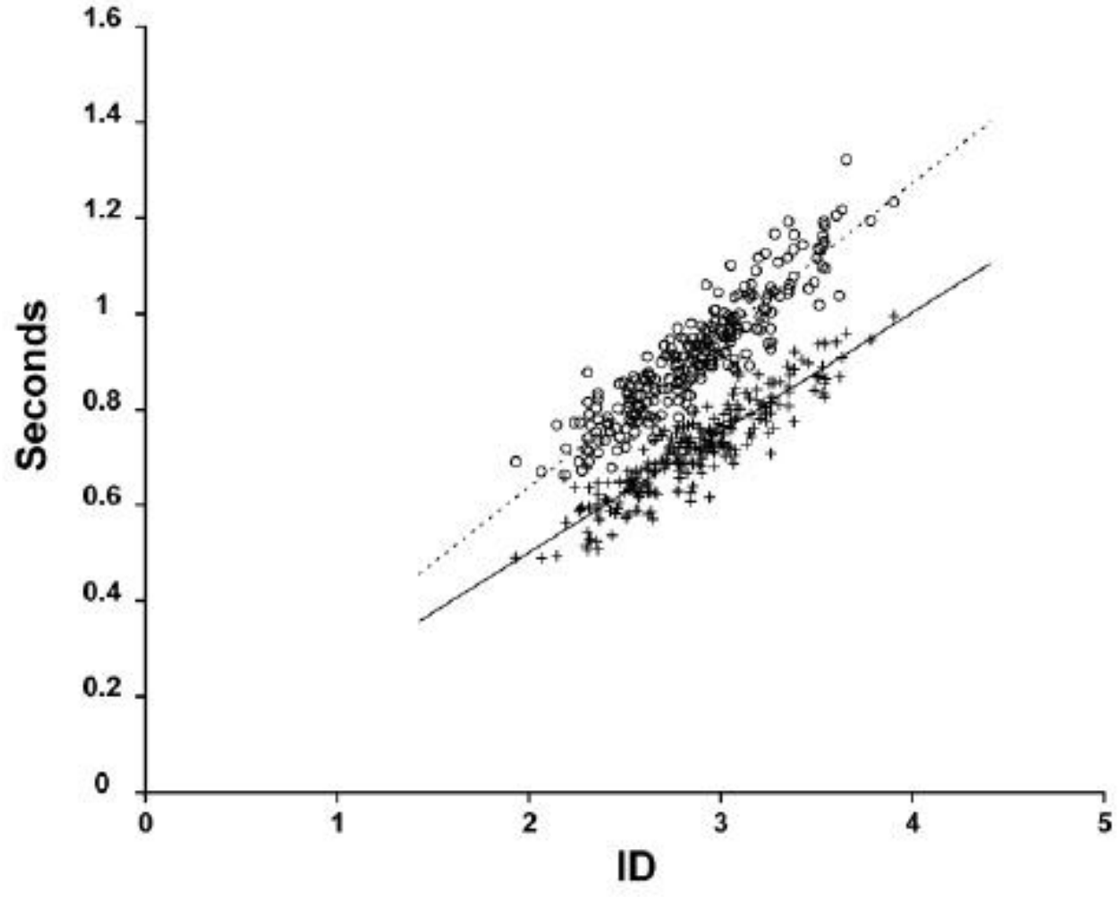
Correlation

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

Multi-variate analysis of variance (MANOVA)

- Use for multiple within-subjects conditions

Repeated measures analysis of variance (RM-ANOVA)

- Use when experiment repeated with each subject (within subjects expt.)
- Complex requirements, not usually applicable, error-prone

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

Mixed Designs

For this experiment, neither between nor within designs are very practical. Why?

A mixed design can use device type (pen, mouse) as a between subjects variable, and the menu task as a within subjects variable.

That is, subjects are split into two groups, A and B.

A subjects use pen only, and B use mouse only. Every subject completes all 4 menu tasks.

MANOVA

MANOVA is Multivariate Analysis of Variance. It is a good general purpose approach to mixed designs like we just described. To run a MANOVA analysis, you build data tables like this:

	menu (within) conditions				device (between)
Subject	A	B	C	D	condition
1	2.1	3.1	4.4	2.7	1
2	1.9	4.2	3.5	3.8	1
:					:
21	2.1	3.3	4.2	2.8	2
22	1.7	4.5	3.2	1.9	2
	X				G

MANOVA

To run a MANOVA analysis in Matlab, call

Manova1(X, G, s);

Where s is the desired significance level (say 0.05).

Matlab returns a verdict on the Null hypothesis – either accept or reject, and a variety of other statistical data.

Discounting

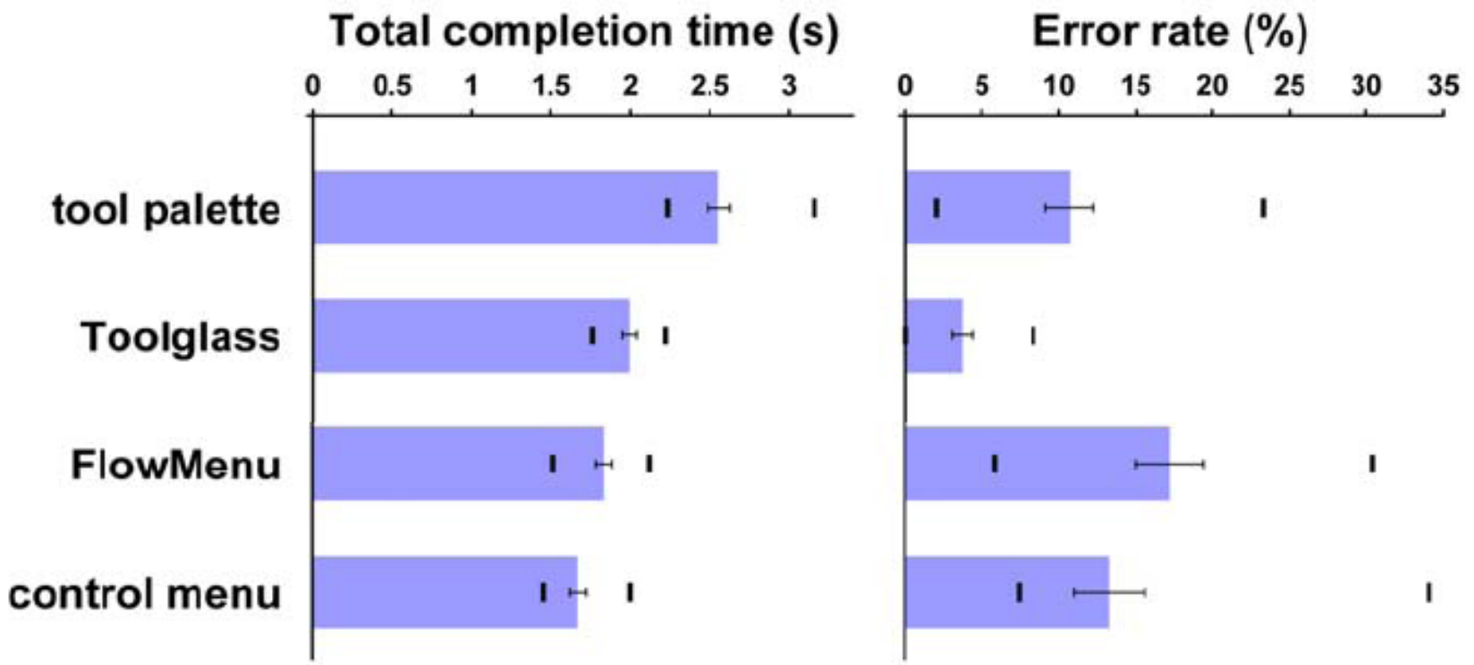
The p-value is the probability of the experiment succeeding by random chance. Thresholds (significance levels) of 0.01 to 0.05 are often used.

But notice that at p = 0.05, if you repeated the experiment 20 times, you have a good chance of "confirming" the hypothesis even if it is false.

Suppose instead you publish a paper with 20 experiments, each of which "proves" its hypothesis at p-value 0.05 ?

If a paper contains many results (e.g. paired t-tests) , the p-values should be decreased to keep the overall significance of the paper low.

Menu Selection Example



- 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

Draw Conclusions

What is the scope of the finding?

- Does the experiment reflect real use?
  - External validity
  - Ecological validity
- Are there other parameters at play?
  - Internal validity

Summary

Quantitative evaluations

- Repeatable, reliable evaluation of interface elements
- To control properly, usually limited to low-level issues
  - Menu selection method A faster than method B

Pros/Cons

- Objective measurements
  - Good internal validity → repeatability
- But, real-world implications may be difficult to foresee
- Significant results doesn't imply real-world importance
  - 3.05s versus 3.00s for menu selection

Project Presentations

Monday:

Group: Orquestra
Group: Lucky Seven
Group: Phi-tus
Group: BuTtErFlY
Group: !Xobile

Project Presentations

Wednesday:

Group: 1 3 3 7
Group: TBD
Group: Group Ate
Group: 4

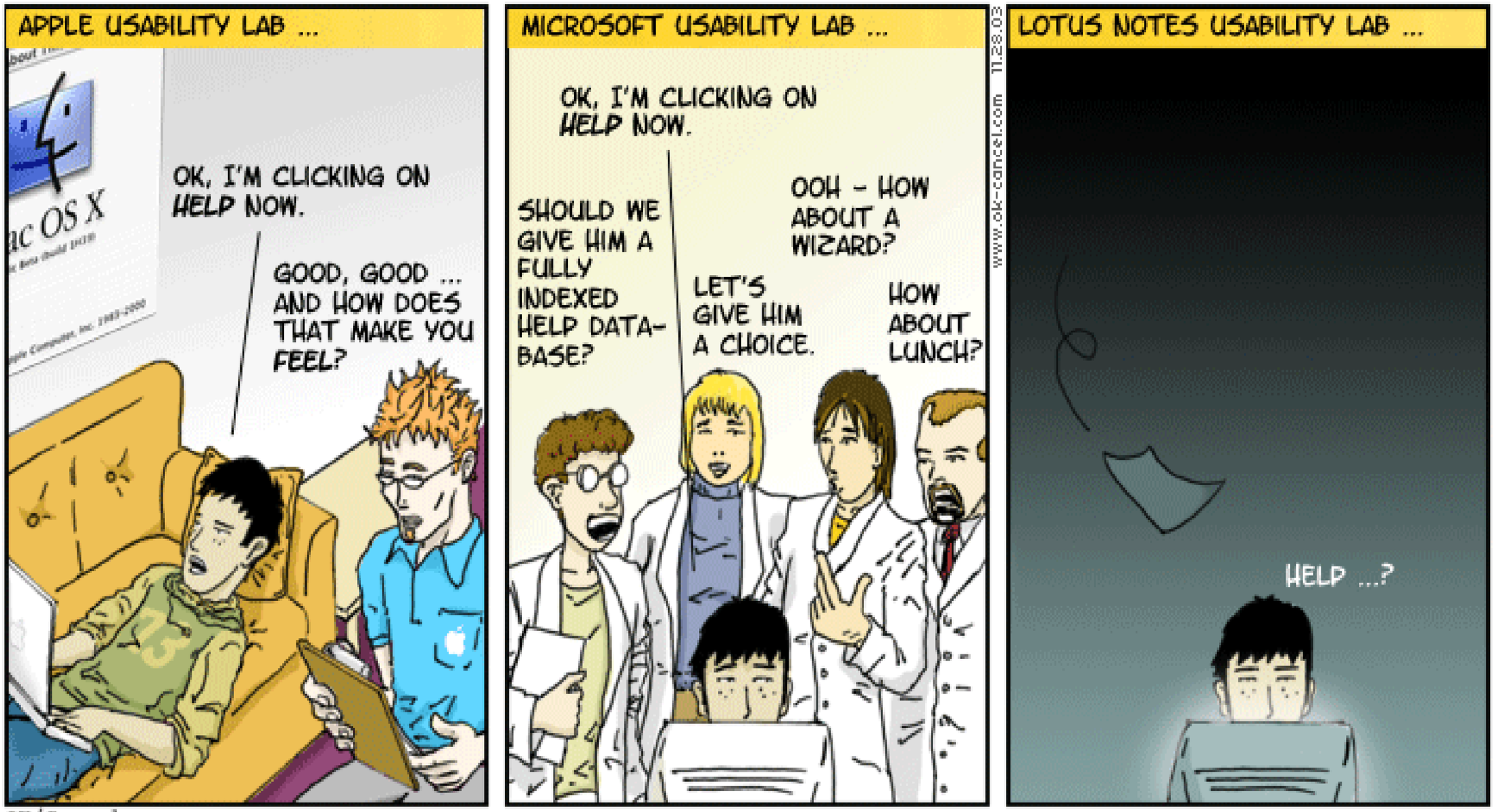
Project Presentations

Aim for 10 minutes for presentation and 3 minutes for questions.

2 or 3 presenters – every group member should present next week or during final presentations.

Test hardware! Make sure your laptop operates correctly with the projector. Bring the presentation on flash drive.

Errors and Help



Errors and Help

- Exercise (2 minutes):
  - List 4 different errors that can occur in your group's interface

Types of User Errors

- **Slips** are errors where a user formulated the correct goal, but carried it out incorrectly
- **Mistakes** are a failure to formulate the right intention

Two Types of Mistakes

- **Mistakes** generally fall into two categories:
  - **Knowledge-based mistake:** Incorrect decision/action because of a failure to understand the situation.
  - **Rule-based mistake:** Understand the situation, but making a wrong decision.

Types of User Errors

- **Slips** are errors where a user formulated the correct goal, but carried it out incorrectly
- **Mistakes** are a failure to formulate the right intention
- **Lapses:** Failure to carry out an action. (Often when part of a sequence is skipped)
- **Mode errors:** Action is correct in one mode of action, but wrong in another.

- The difference matters because:
  - The method used to fix the user interface is different: how?

Possible Causes

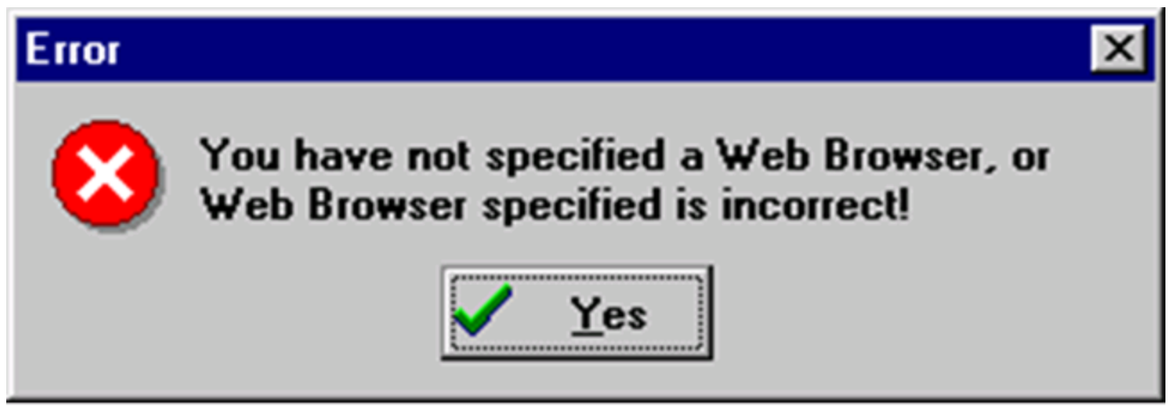
- Incorrect cause and effect
- Inadequate background to understand the information
- Unclear understanding of system status
- Misjudging information importance

Preventing Errors

- Exercise (2 minutes):
  - For each of your errors, classify it as a slip, mistake, lapse or mode error and design a way to fix the error

- What is the best way to prevent errors?

System Errors



- Write in the user's language
  - "winword.exe" caused a segmentation fault at #F34EA01.
  - You need to know the understand the users to do this
- Precisely indicate the problem
- Constructively suggest a solution

Helping Users Learn

- How do we help users learn our system so they make fewer errors?

Help (doesn't)

- Extra feature that can confuse users
- Spreading expensive jam onto stale toast isn't going to make it taste better

- In a 1987 study of 52,576 help sessions:
  - 23% of all requests found no help
  - 36% of people who found help reported the help was useful (28% of total requests)

Helping Help Help

- People want answers, and want them quickly

- Descriptive questions; "What is this?"
- Procedural questions; "How do I do this?"
- Guidance questions; “What should I do?”
- Interpretive questions; "Why did that happen?"
- Navigational questions; "Where am I?“, “Where is X?”

Types of Help

- F1 help
- …?

Cost of Help

- What is the most expensive form of help?
  - Asking a friend

- What is the least expensive form of help?
  - A computer interface that doesn't need help

Experts and Beginners

- Who are they?
- How do we design for them?

Beginners

- User Description
  - System knowledge:
    - None
  - Domain Knowledge:
    - Unknown
  - Proficiency:
    - Low

How Beginners will Behave

- Few tasks
- Many errors
- Dependence on help (not just heavyweight help)
- Limited use of options or alternatives

Supporting Beginners

- Few options
- Visible help
- At most one task per screen
- Wizards
- Provide acquisition facilities
  - Highly visible
  - Aesthetically pleasing
  - Concentrate on ordinary, standard, typical tasks

Experts

- User Description
  - System knowledge:
    - High
  - Domain Knowledge:
    - High
  - Proficiency:
    - High

How Experts will Behave

- Many tasks
- Few errors
- Little use for Help
- Idiosyncratic style of interaction
- High use of options or alternatives
- Primary concern is efficiency and productivity

Supporting Experts

- Efficient Interaction
- Fast
- Many tasks per screen
- Provide production facilities
  - Conventional and Familiar techniques to support expert use
    - Ctrl+x, ctrl+c, ctrl+v
  - Uncluttered, customizable workspace
  - Simple icons on toolbars and dockable toolbars
  - Features that rely on user's memory rather than visibility

Unix-style Command Line

- How many people are beginners?
  - % cp ~/Desktop/myhouse.png ~/Desktop/pictures/myhouse.png

- How many people are experts?
  - % for file in $(find . -name *.png -print) ; do convert \
    -size 800x800 ${file} -resize 800x800 ${file//.png}-small.png \
    ; done

- Most users of software are “perpetual intermediates” or “improving intermediates”

How Intermediates will Behave

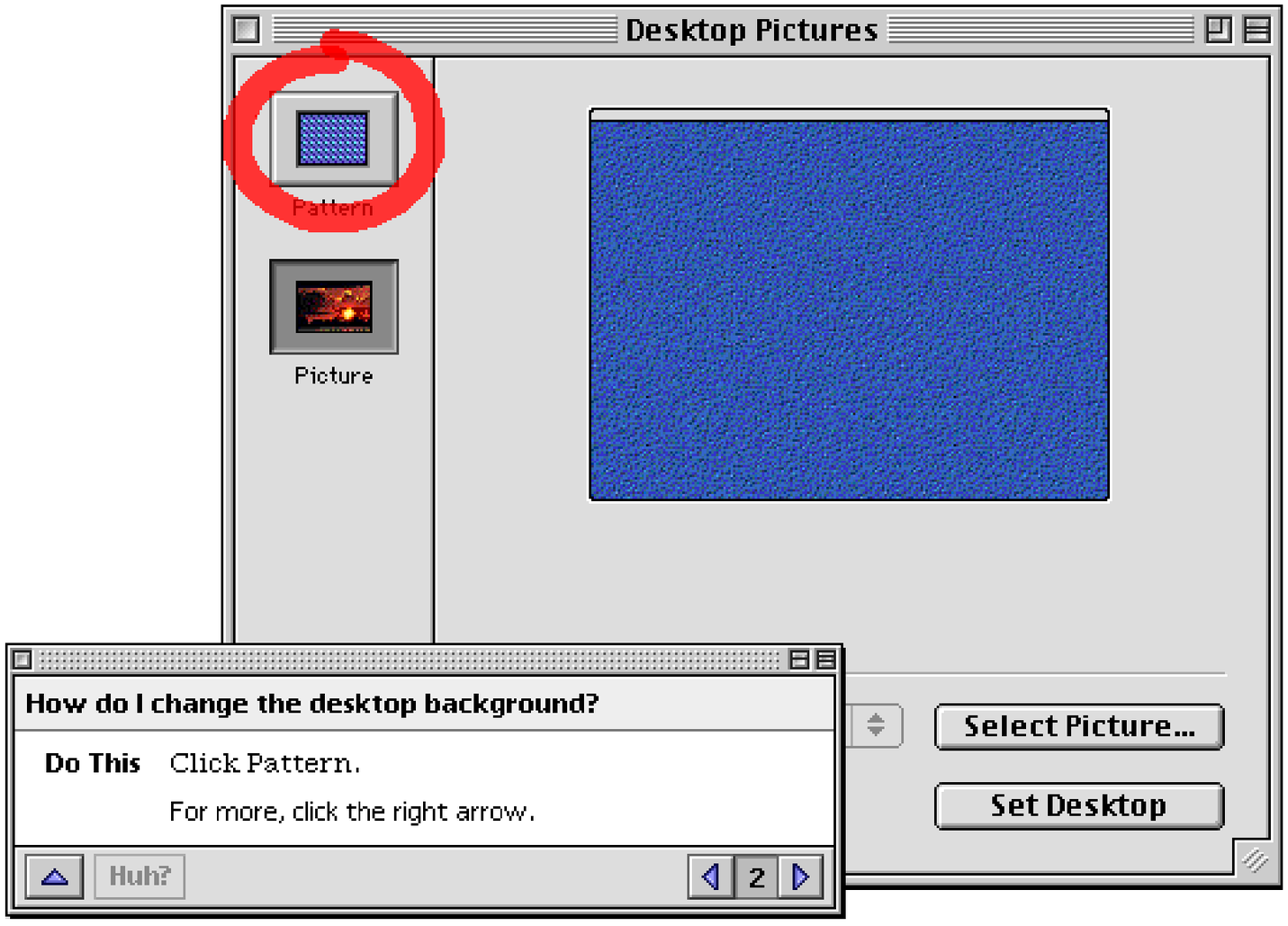
- Expanding number of tasks
- System limitations become frustrating
- Intermittent need for help
- More extensive experimentation
- Evolving and changing patterns of interaction

Interfaces for Intermediates

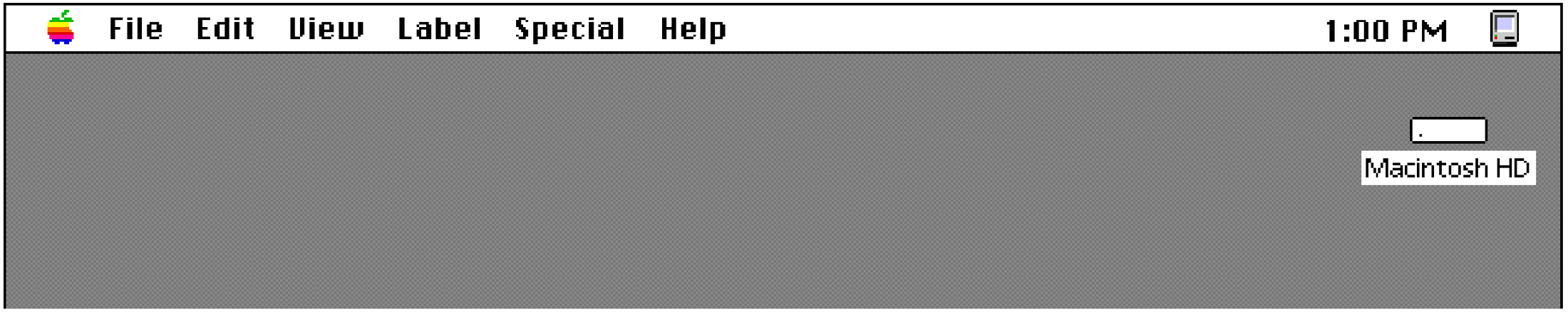
- Allow exploration through interaction
- Show alternate mechanisms to perform tasks
- Provide transitional facilities
  - Visible shortcuts
  - Customizable interface

Apple Help

Apple Guide

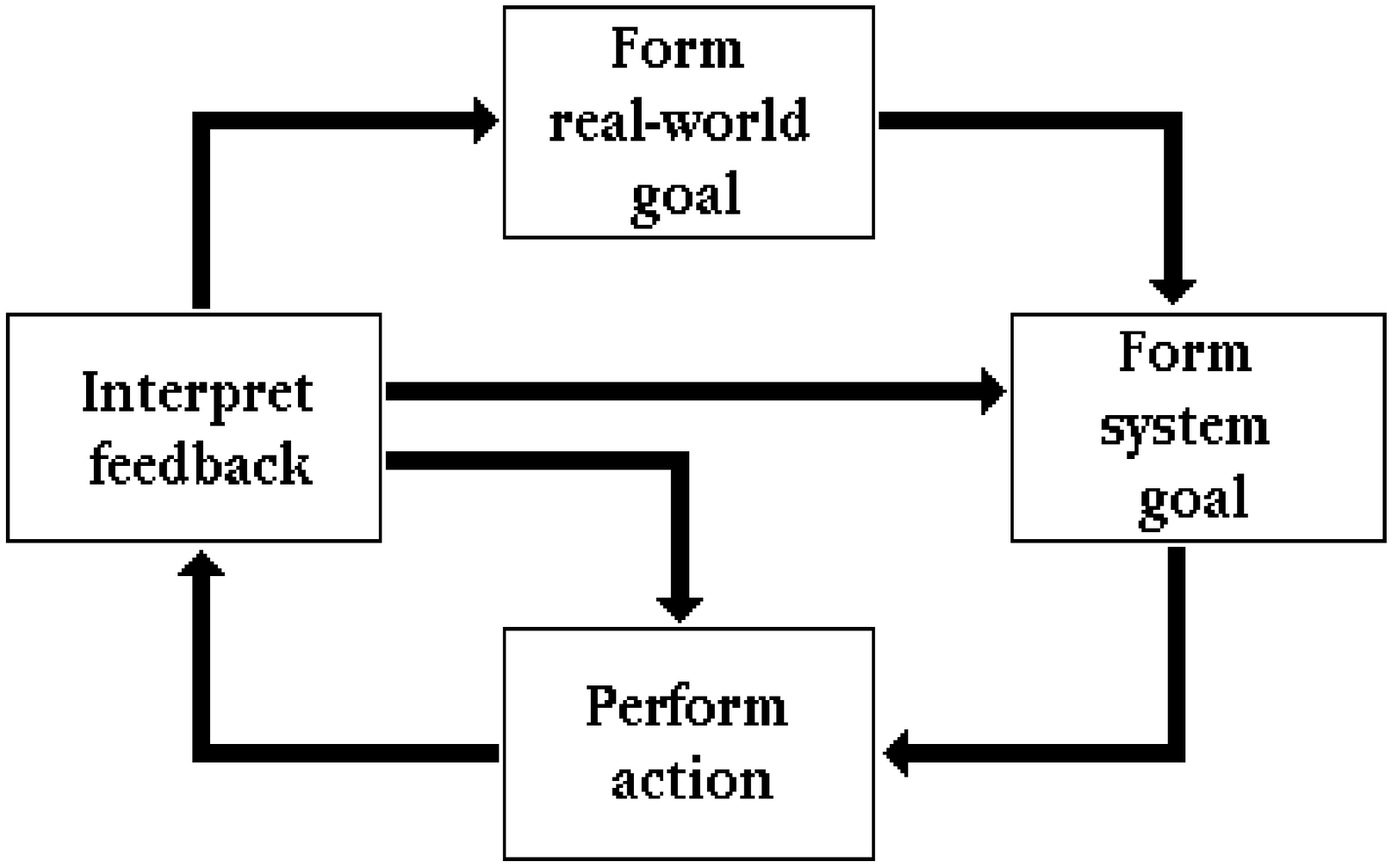

Desktop Pictures
Pattern
Picture
Select Picture...
Set Desktop
How do I change the desktop background?
Do This
Click Pattern.
For more, click the right arrow.
Huh?
2

Discoverability

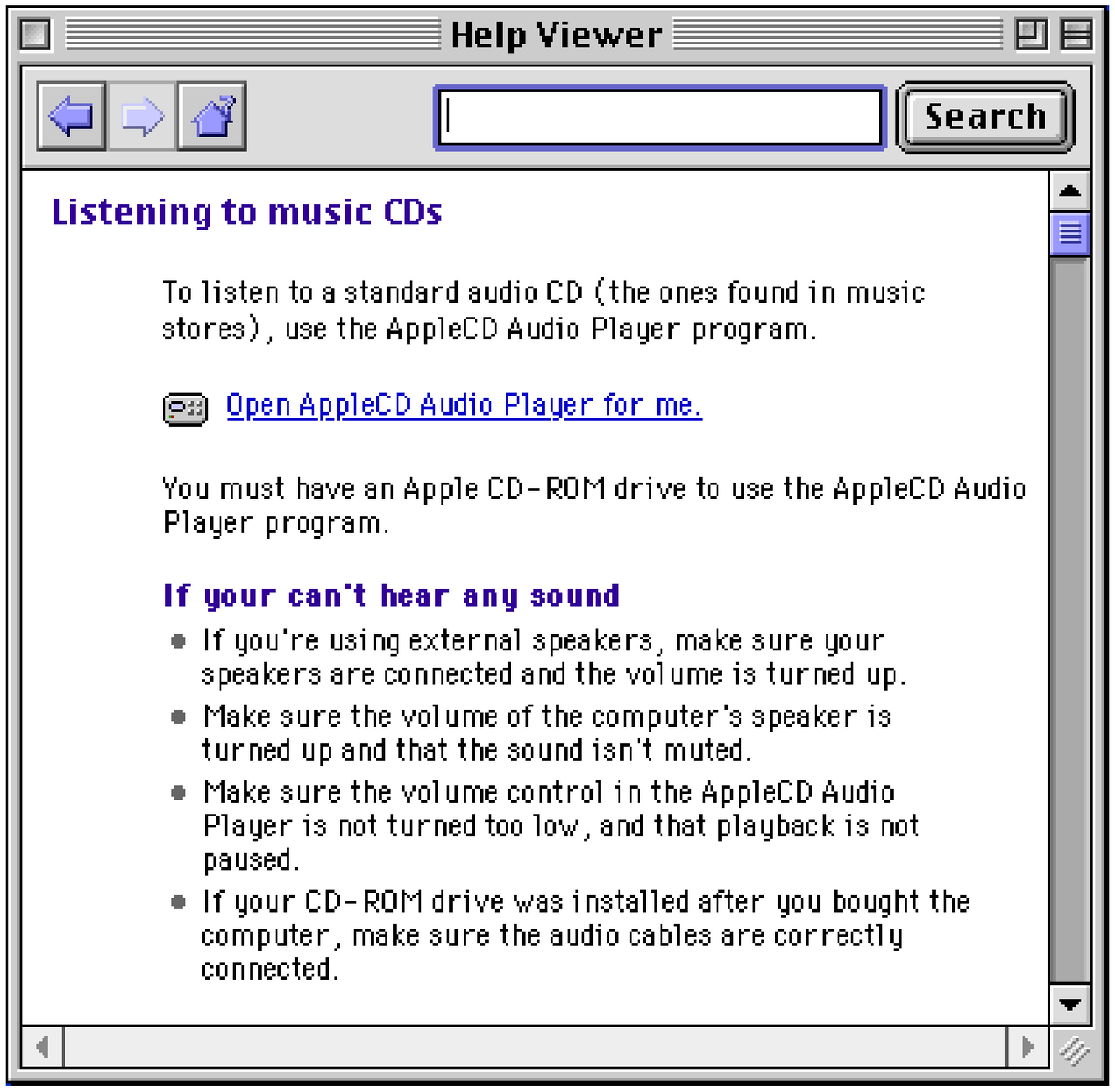


Central Access

Define Tasks Broadly



Write Minimally



Best practices in Game Help?

- Mission objectives
- Where things are
- Tool tips + behaviors
- Encourage exploration

Conclusion

- Wrapped up quantitative testing
- Descriptive statistics
- Types of test, assumptions
- Help – targeting users
- Help – Apple help principles

- Next time: Interactive Prototype presentations
  - You will be reviewing other groups to give feedback