Lecture 12 Evaluation - 2

Xiaojun Bi
Stony Brook University
xiaojun@cs.stonybrook.edu

Previous Lecture

Qualitative vs. Quantitative Evaluation

Heuristic Evaluation

Developed by Jakob Nielsen (1994)

Can be performed on working UI or sketches

Small set (3-5) of evaluators (experts) examine UI

- Check compliance with usability heuristics
- Different evaluators will find different problems
- Evaluators only communicate afterwards to aggregate findings
- Use violations to redesign/fix problems

Heuristics

H2-1: Visibility of system status

H2-2: Match system and real world

H2-3: User control and freedom

H2-4: Consistency and standards

H2-5: Error prevention

H2-6: Recognition rather than recall

H2-7: Flexibility and efficiency of use

H2-8: Aesthetic and minimalist design

H2-9: Help users recognize, diagnose and recover from errors

H2-10: Help and documentation

Agenda

- Controlled Experiment
  Independent vs. Dependent Variables
  Within Subject vs. Between Subject Design
  Descriptive vs. Inferential Statistics

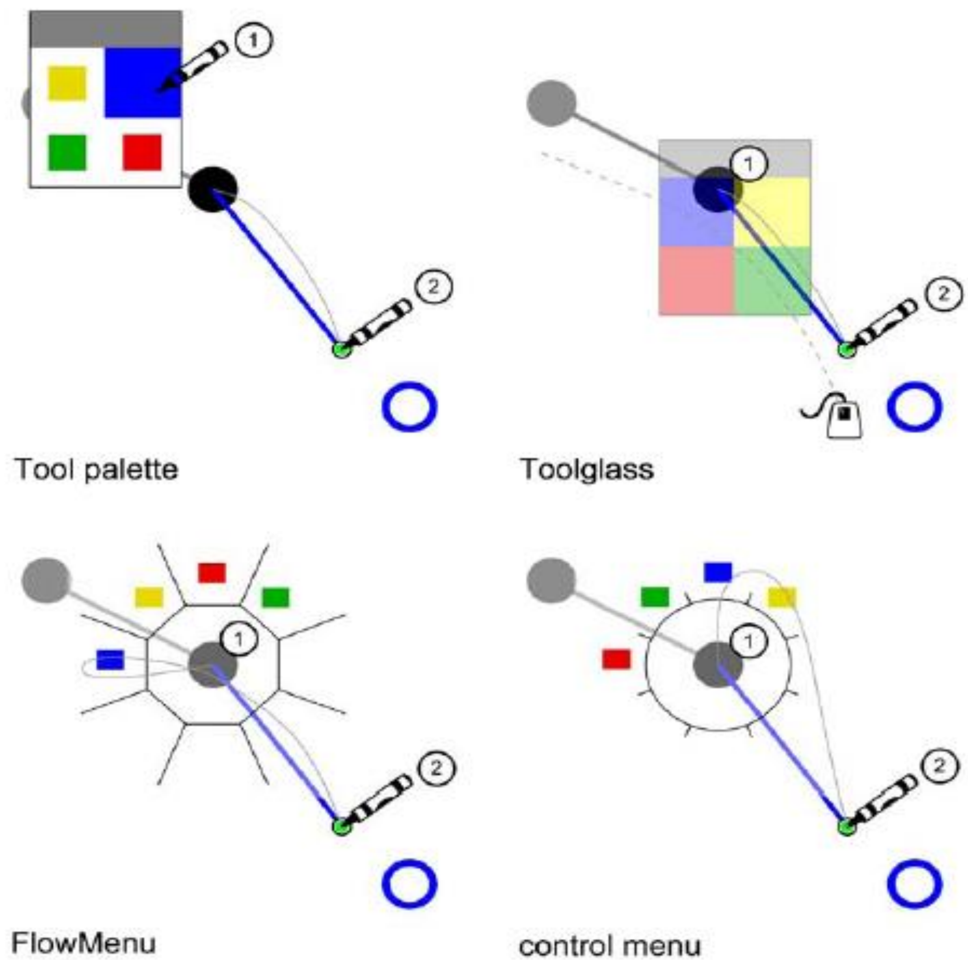
- Interaction Effects

Procedures of Running a Controlled Experiment

Steps in Designing an Experiment

1. State a lucid, testable hypothesis
2. Identify variables (independent, dependent)
3. Design the experimental protocol
4. Run pilot studies
5. Run the experiment
6. Perform statistical analysis
7. Draw conclusions

Example: Menu Selection



François Guimbretière, Andrew Martin, and Terry Winograd, Benefits of Merging Command Selection and Direct Manipulation. Transactions on Human-Computer Interaction, 12(3), pp 460 – 476, 2005

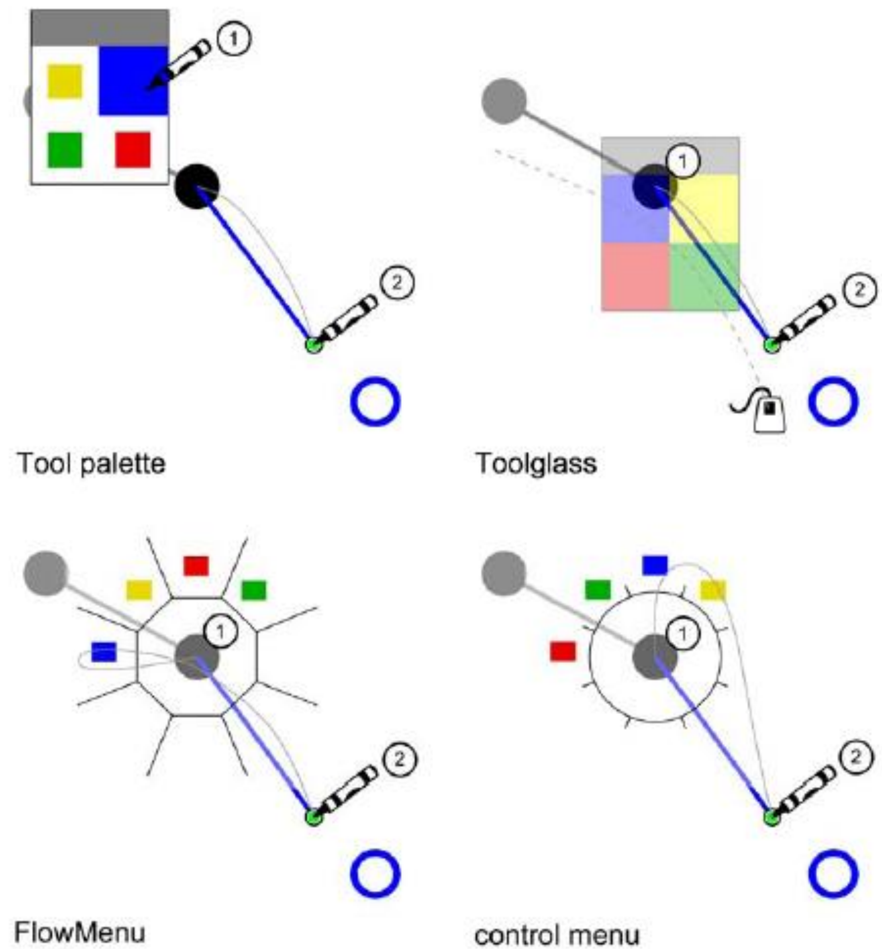
Steps in Designing an Experiment

1. **State a lucid, testable hypothesis**
2. Identify variables (independent, dependent)
3. Design the experimental protocol
4. Run pilot studies
5. Run the experiment
6. Perform statistical analysis
7. Draw conclusions

Lucid, Testable Hypothesis

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

Other hypotheses?



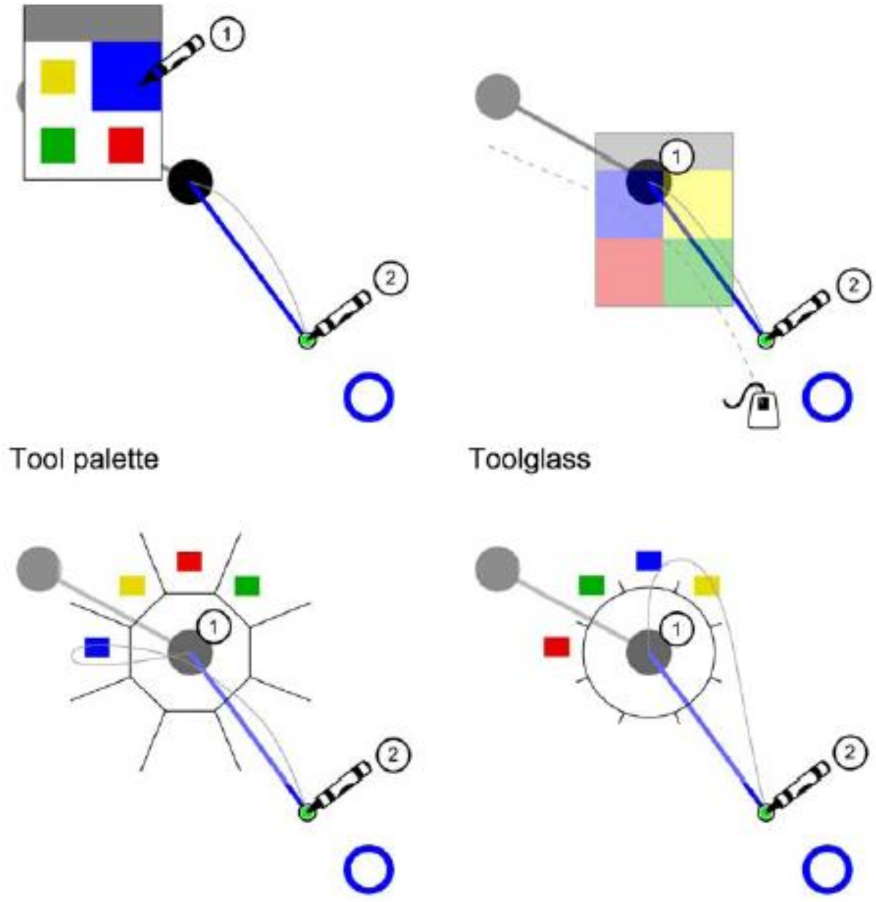
Steps in Designing an Experiment

1. State a lucid, testable hypothesis
2. **Identify variables (independent, dependent)**
3. Design the experimental protocol
4. Run pilot studies
5. Run the experiment
6. Perform statistical analysis
7. Draw conclusions

Variables

Independent variables (IV)

Dependent variables (DV)



Tool palette

Toolglass

FlowMenu

control menu

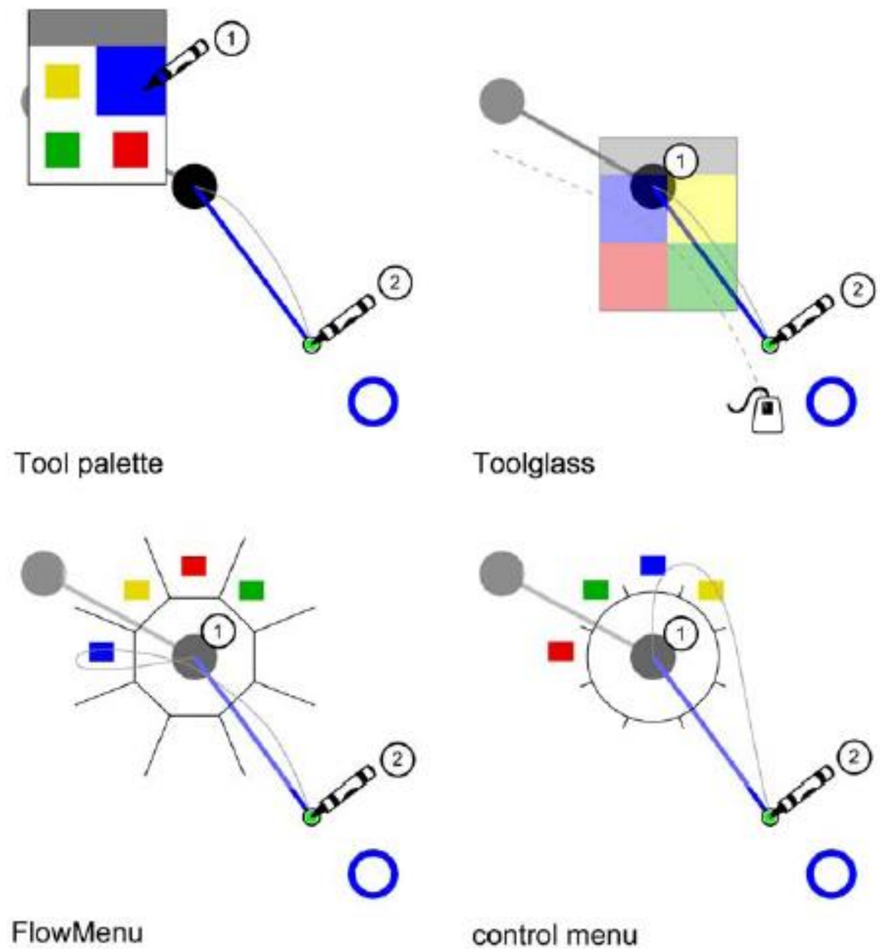
Variables

Independent variables (IV)

- Menu type (4 choices)

Dependent variables (DV)

- Time
- Error rate
- User satisfaction



Internal and External Validities

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

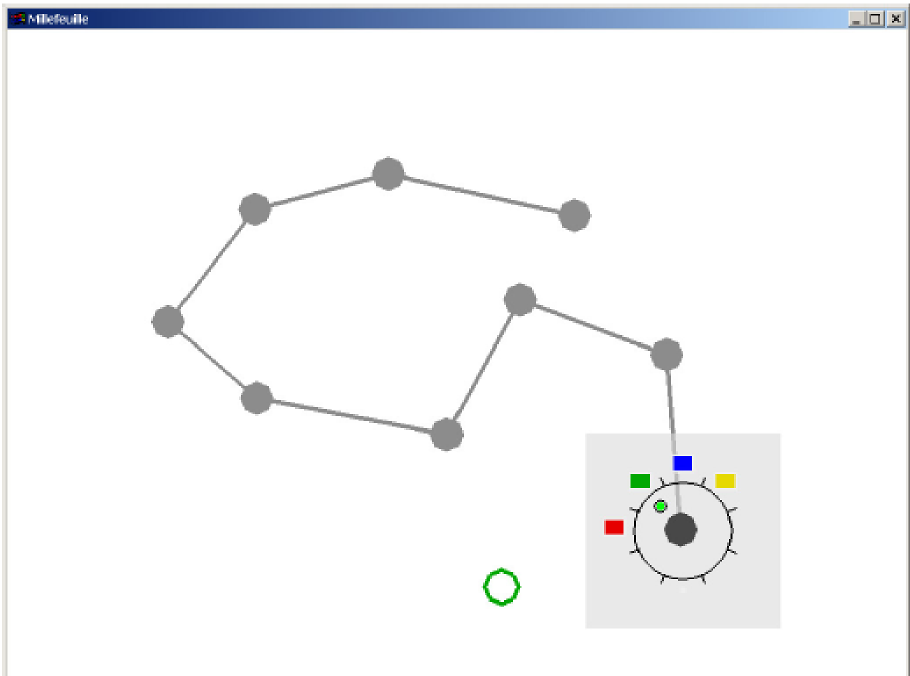
Steps in Designing an Experiment

1. State a lucid, testable hypothesis
2. Identify variables (independent, dependent)
3. **Design the experimental protocol**
4. Run pilot studies
5. Run the experiment
6. Perform statistical analysis
7. Draw conclusions

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.

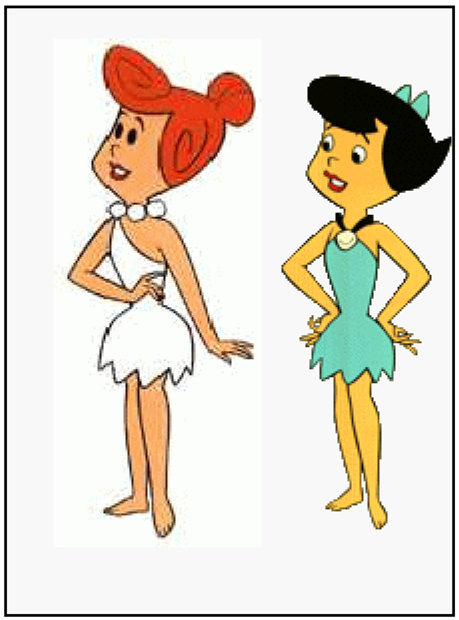
Number of Conditions

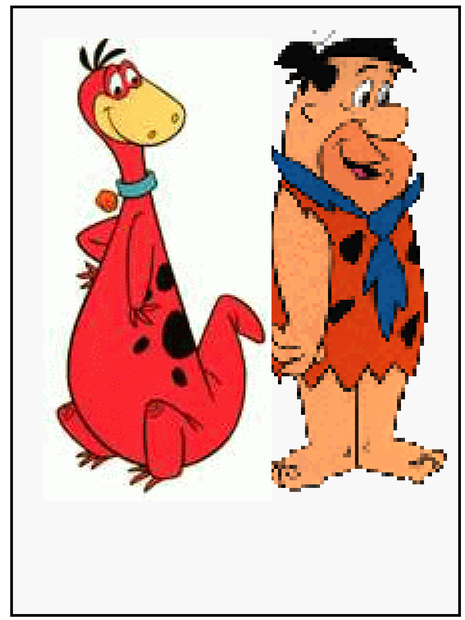
- Tool Palette
- Tool Glass
- Flow Menu
- Control Menu

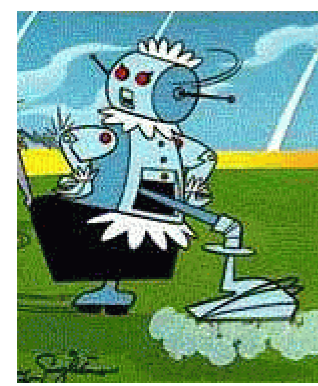
Between Subjects Design

Wilma and Betty use one interface

Dino and Fred use the other

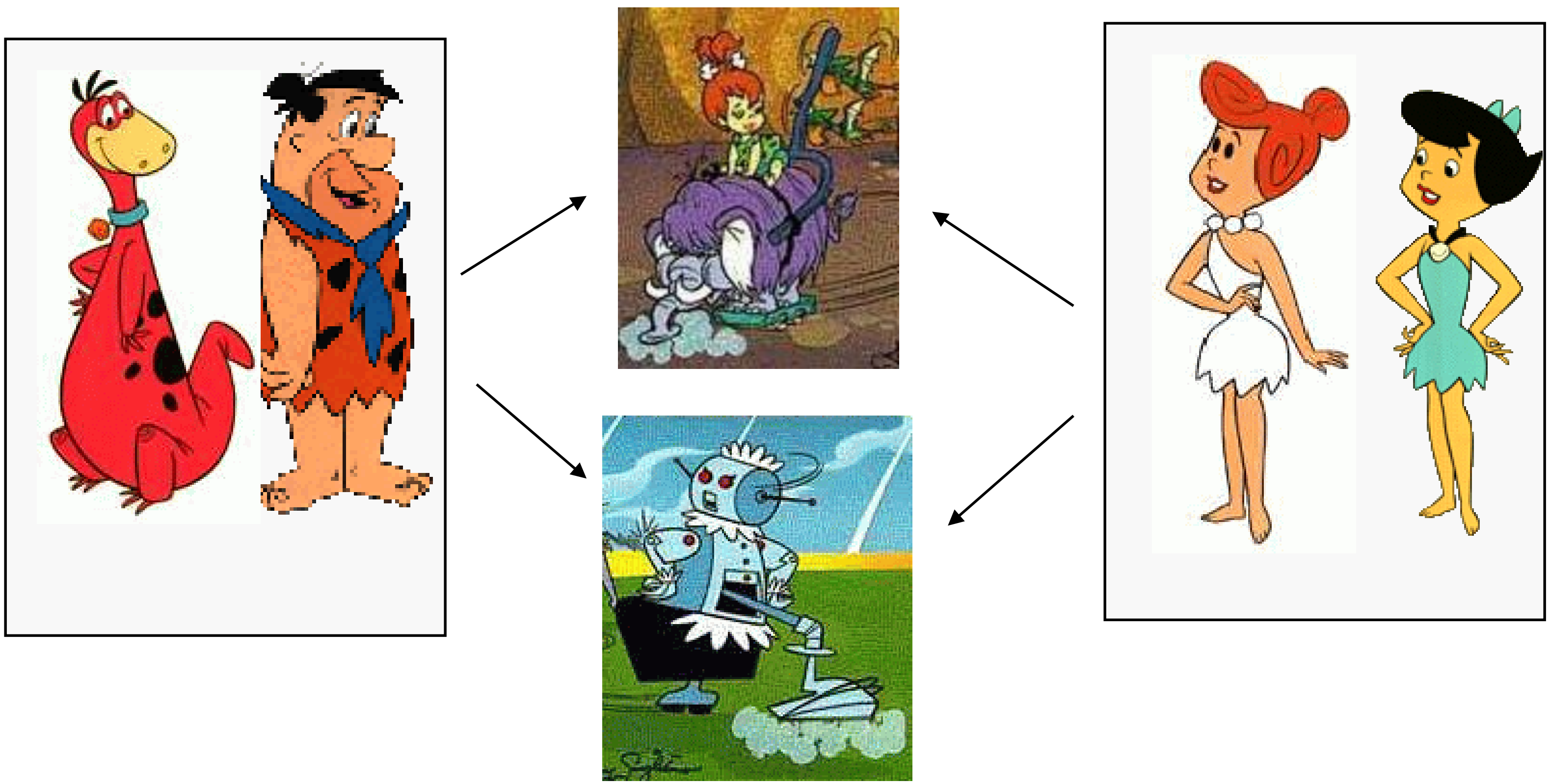






Within Subjects Design

Everyone uses both interfaces



Between vs. Within Subjects

Between subjects

- Each participant uses one condition
  - + Can collect more data for a given condition
  - + Avoid carry-over learning effects
  - - Participants cannot compare conditions
  - - 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

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

Steps in Designing an Experiment

1. State a lucid, testable hypothesis
2. Identify variables (independent, dependent)
3. Design the experimental protocol
4. **Run pilot studies**
5. **Run the experiment**
6. Perform statistical analysis
7. Draw conclusions

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

Steps in Designing an Experiment

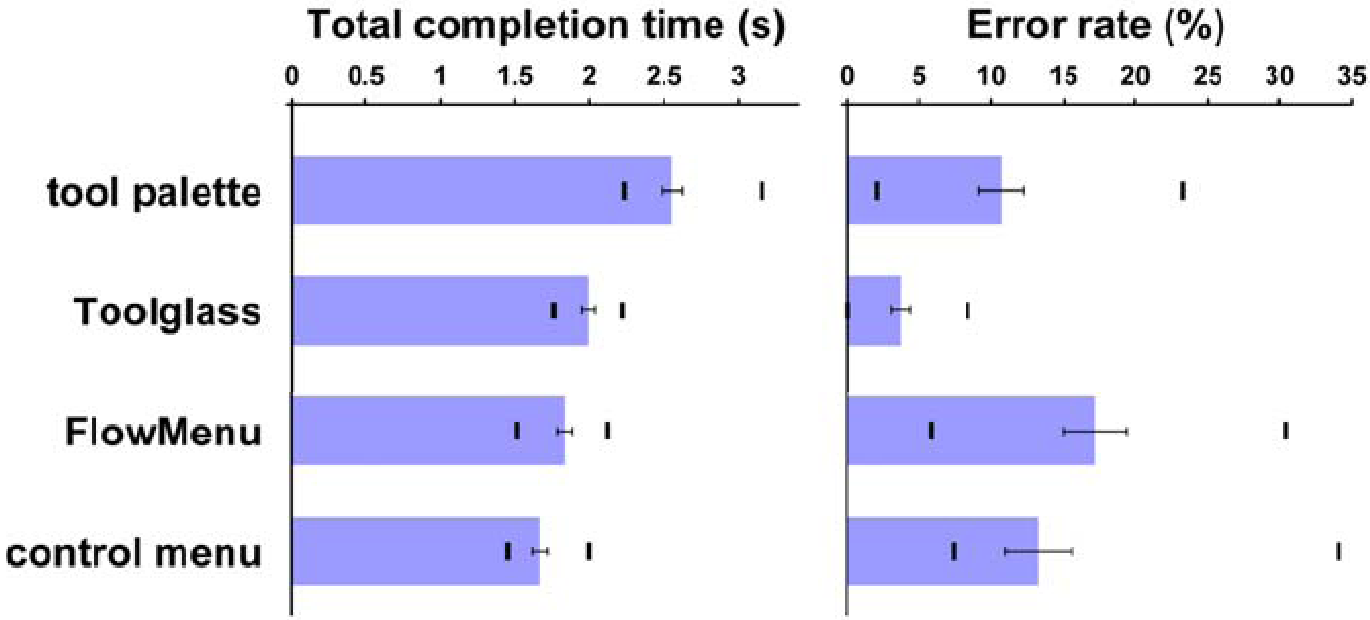
1. State a lucid, testable hypothesis
2. Identify variables (independent, dependent)
3. Design the experimental protocol
4. Run pilot studies
5. Run the experiment
6. **Perform statistical analysis**
7. Draw conclusions

Descriptive Statistics

Results: Statistical Analysis

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

- Mean
- Standard deviation



Inferential Statistics

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)
- ANOVA (1 factor, > 2 levels, multiple factors)

Explaining Psychological Statistics
Barry H. Cohen

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

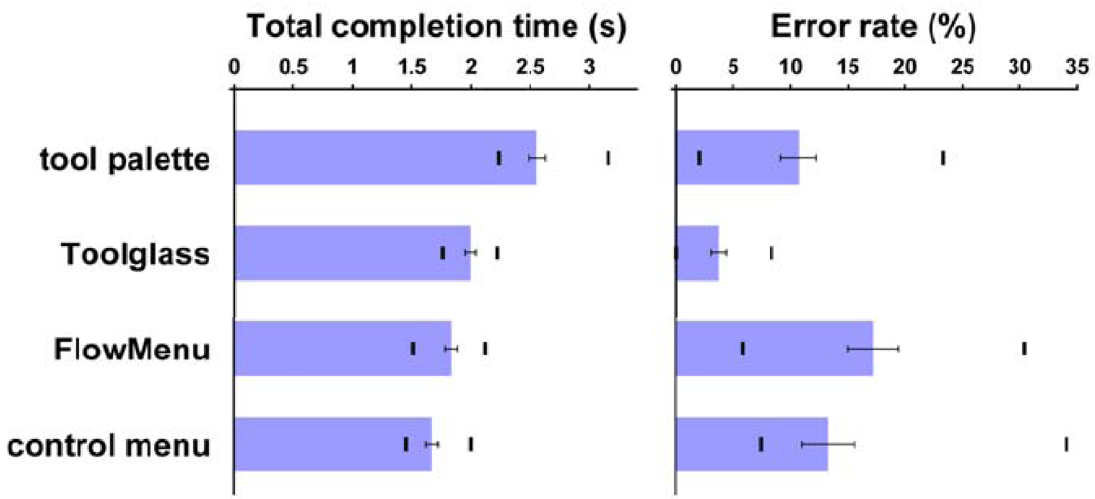
Multi-variate analysis of variance (MANOVA)

- 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



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)

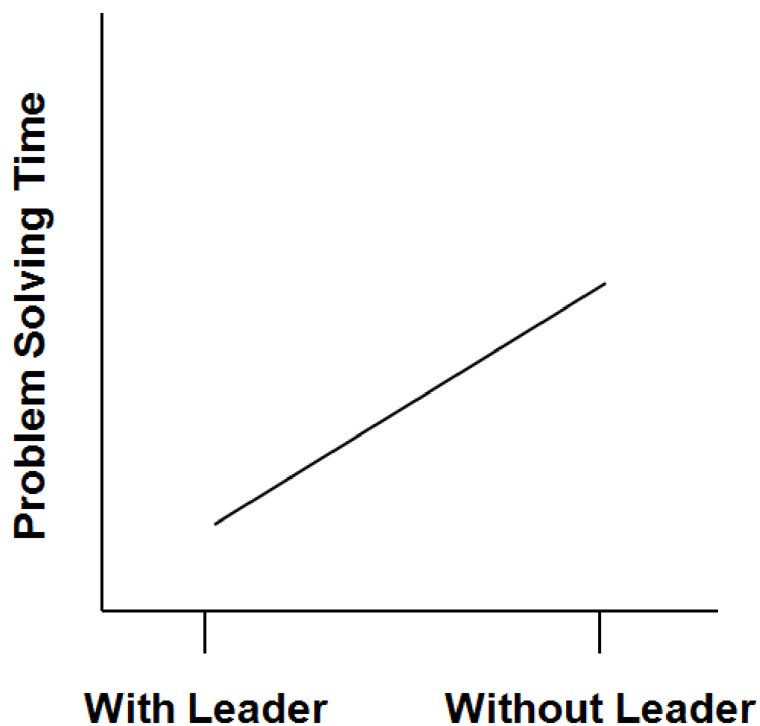
Separate analysis for error rates

Interaction Effects

Example of Interactions

Group problem solving

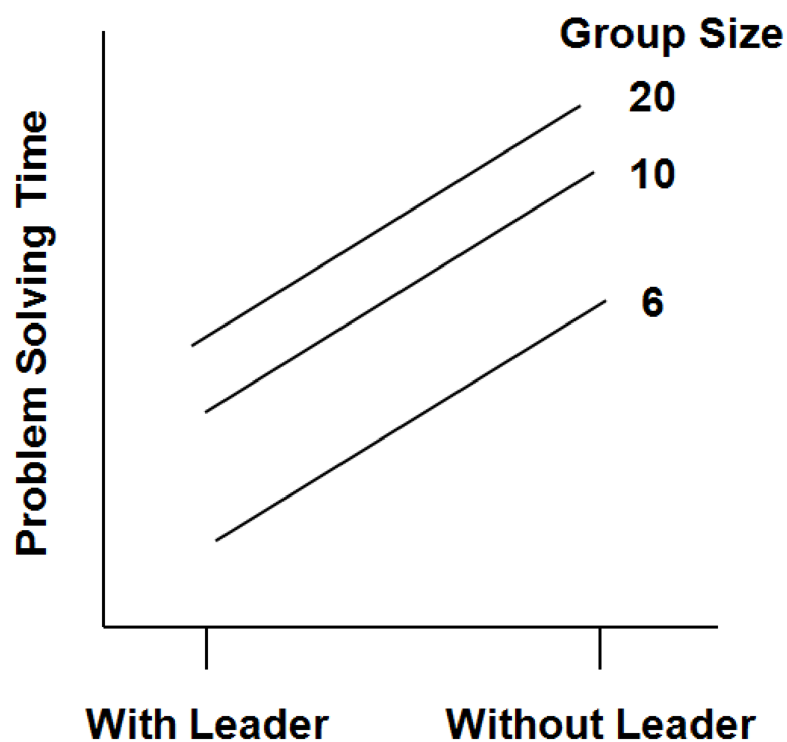
- Independent variable: Leadership



[from Martin 04]

Example of Interactions

Group problem solving
- Independent variable: Leadership
- Independent variable: Group size

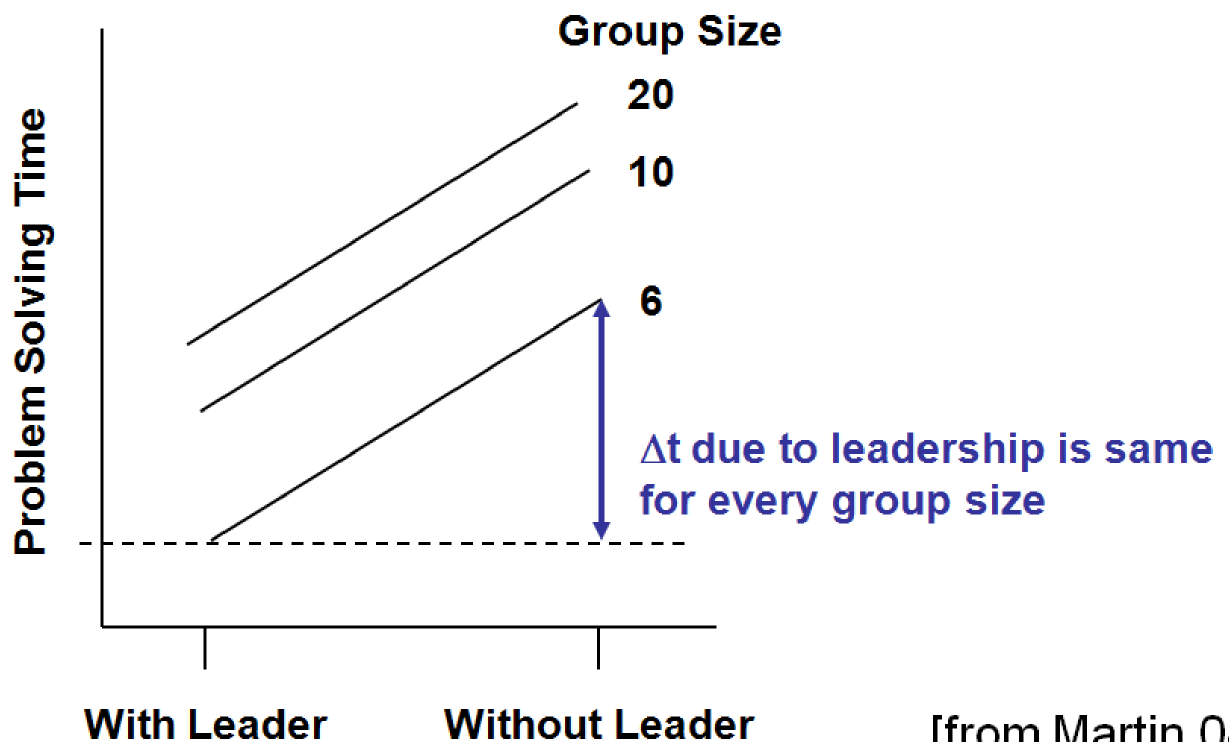


[from Martin 04]

Example of Interactions

Group problem solving

- Change in time due to leadership is same regardless of group size

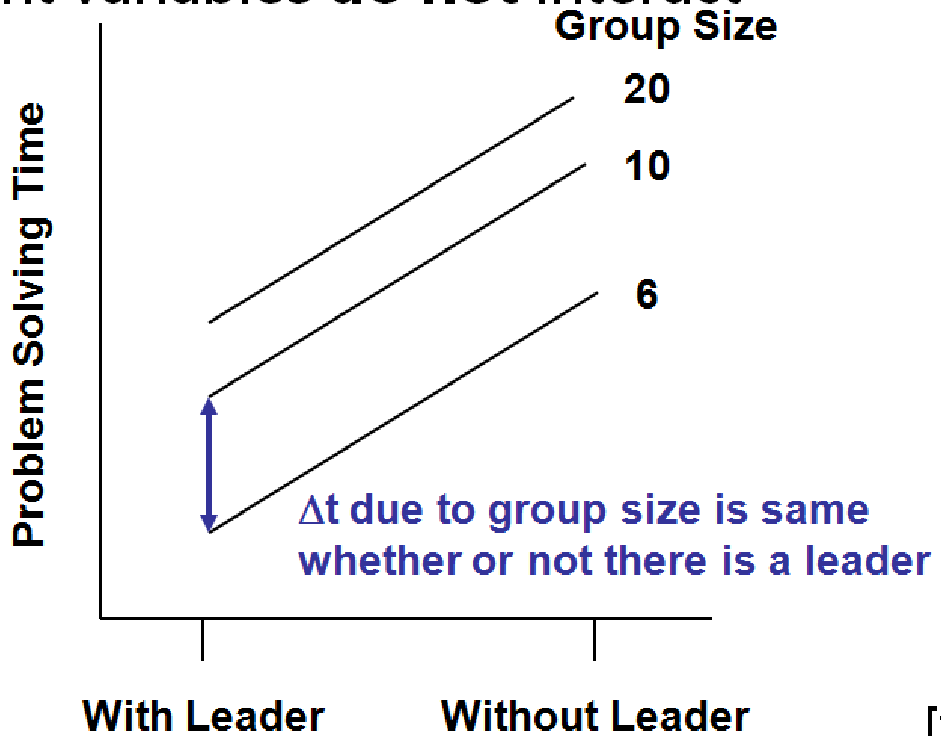


[from Martin 04]

Example of Interactions

Group problem solving

- Change in time due to leadership is same regardless of group size
- Change in time due to group size is same regardless of leadership
- Independent variables **do not** interact

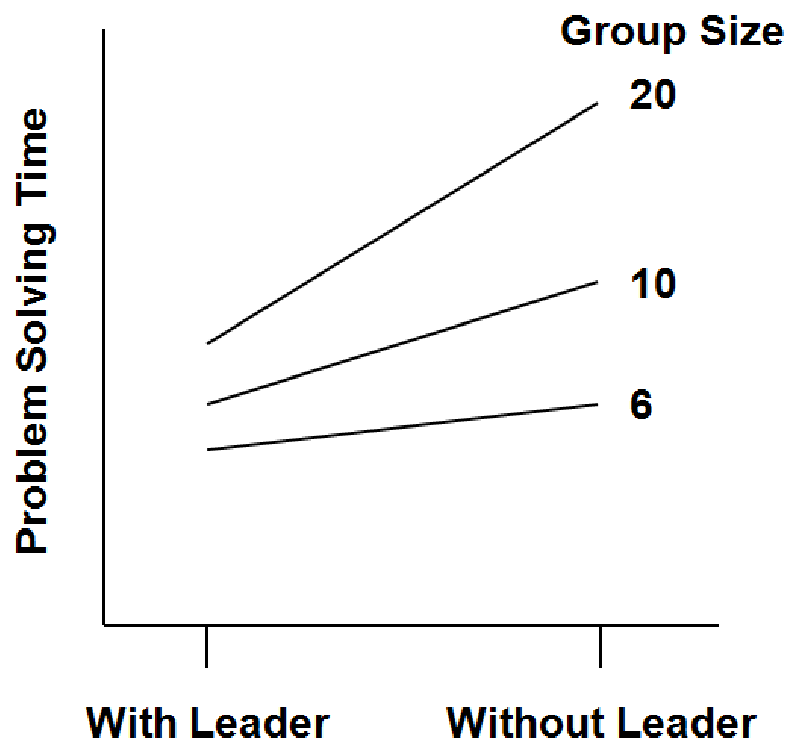


[from Martin 04]

Example of Interactions

Multiple IVs effect DV non-additively

- Change in time due to leadership differs with changes in group size
- Independent variables **do** interact



[from Martin 04]

Summary

- Controlled Experiment

  Independent vs. Dependent Variables

  Within Subject vs. Between Subject Design

  Descriptive vs. Inferential Statistics

- Interaction Effects