CS160: User Interface Design

Midterm Review 3/12/2012

Berkeley
UNIVERSITY OF CALIFORNIA

Due Today

Lo-Fi Prototype Test Report (now)

New Assignment

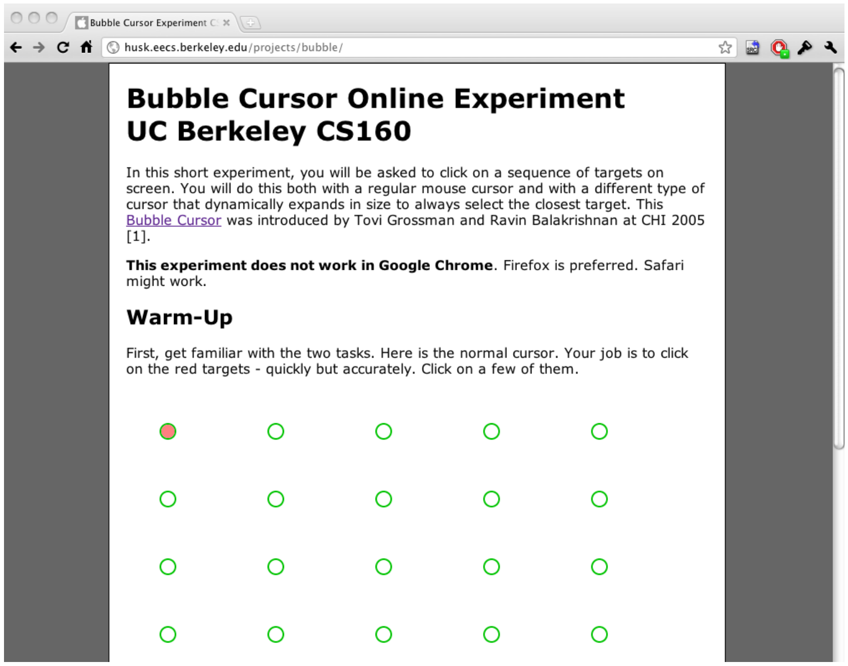
Interactive Prototype: Make your app real!

Functional, interactive app written for Kinect

Due on April 9th

Presentations on April 9, April 11

Data Analysis



Effect Sizes: Time

Normal vs. Bubble cursor at target size 10:
1129ms vs. 796ms: Bubble cursor 30% faster

Normal vs. Bubble cursor at target size 30:
803ms vs. 723ms: Bubble cursor 10% faster

Target size for normal cursor:
1129ms vs 803ms: Larger targets 29% faster

Target size for Bubble cursor:
796ms vs. 723ms: Larger targets 9% faster

Effect Sizes: Error

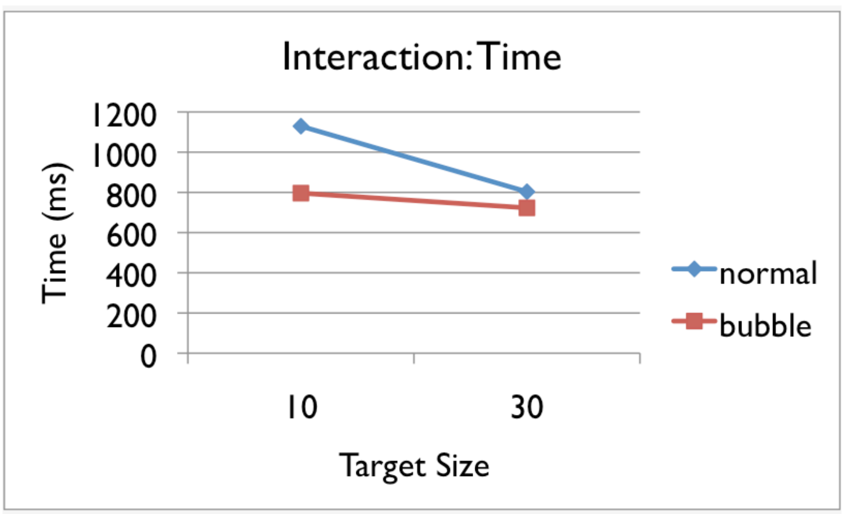
Normal vs. Bubble cursor, target size 10:

1.67 vs. 0.24 Errors per 20 trials: 85% fewer errors! (6.95x)

Normal vs. Bubble cursor, target size 30:

0.90 vs. 0.02 Errors per 20 trials: 98% fewer errors!

Interactions



Interactions

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

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

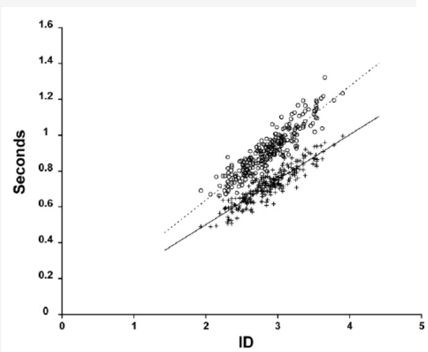
Regression

Compute the "best fit"

linear

logistic

. . .



ANOVA

Single factor analysis of variance (ANOVA)
Compare means 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?

Our Example (Time)

Two-Way ANOVA (Cursor, Size) for **time**:

Main effect for **cursor**

$F(1,1696) = 264.1$, $p<0.001$ is statistically significant

Main effect for **size**

$F(1,1696)=246.7$, $p<0.001$ is statistically significant

Interaction **cursor x size**

$F(1,1696)=92.2$, $p<0.001$ is statistically significant

Our Example (Time)

Still need to run pairwise T-tests

Bubble fixed, Size varying

- $T(858)=22.683, p<0.001$
- Time at size 10 significantly differs from time at size 30

Normal fixed, Size varying

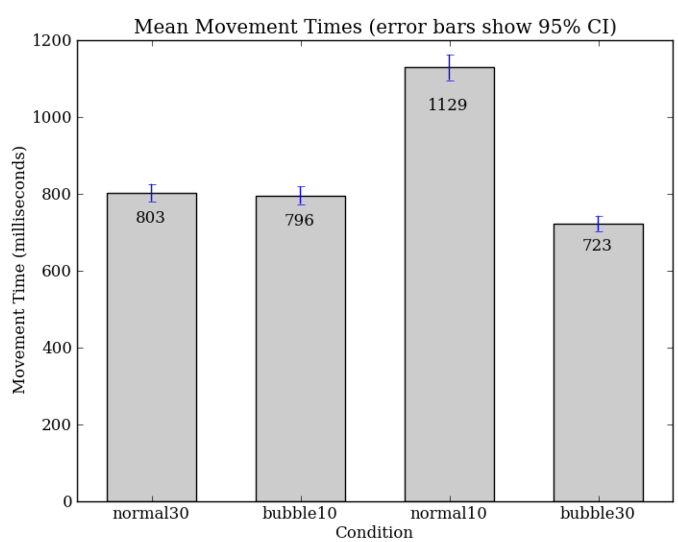
- $T(838)=252.95, p<0.001$
- Time at size 10 significantly differs from time at size 30

Size10 fixed, Cursor varying

- $T(838)=264.72, p<0.001$
- Time with bubble significantly differs from time with normal

Size30 fixed, Cursor varying

- $T(858)=27.11, p<0.001$
- Time with bubble significantly differs from time with normal



Our Example (Errors)

Two-Way ANOVA (Cursor, Size) for **errors**:

Main effect for **cursor**

F(1,81) = 21.0, p<0.001 is statistically significant.

Main effect for **size**

F(1,81)=246.7, **p=0.11** is **not** statistically significant.

Interaction **cursor x size**

F(1,81)=92.2, **p=0.06** is **not** statistically significant.

Our Example (Errors)

Still need to run pairwise T-tests

Bubble fixed, Size varying

- T(41)=0.19, p=0.66
- Errors at size 10 **not** significantly different from errors at size 30

Normal fixed, Size varying

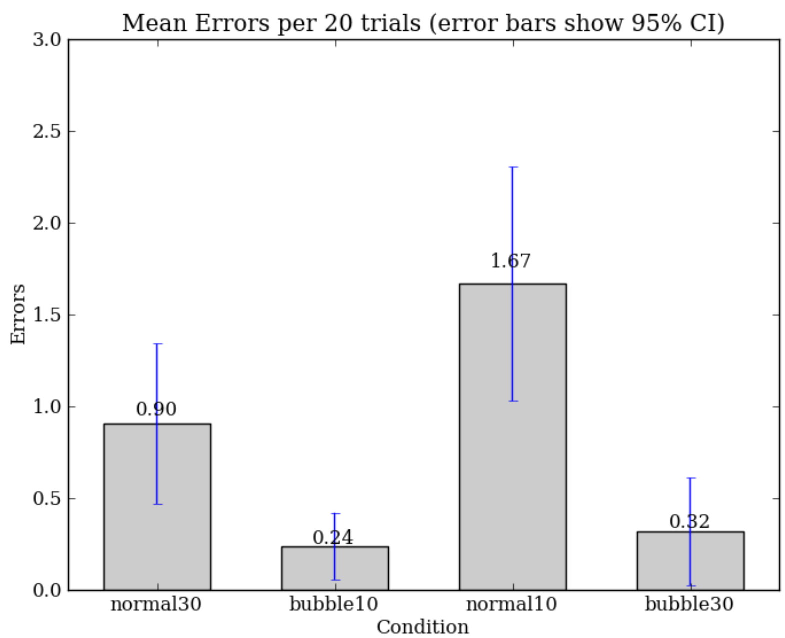
- T(40)=3.56, p=0.066
- Errors at size 10 **not** significantly different from errors at size 30

Size10 fixed, Cursor varying

- T(40)=16.98, p<0.001
- Errors with bubble significantly differs from errors with normal

Size30 fixed, Cursor varying

- T(41)=4.65, p=0.037
- Errors with bubble significantly differs from errors with normal



What does p>0.05 mean?

No statistically significant difference (at 5% level)

Are the two conditions thus equivalent?

NO! We DID observe differences.

But can't be sure they are not due to chance.

Confidence Intervals

95% Confidence Interval: The range of values in which we're 95% sure the true population mean falls.

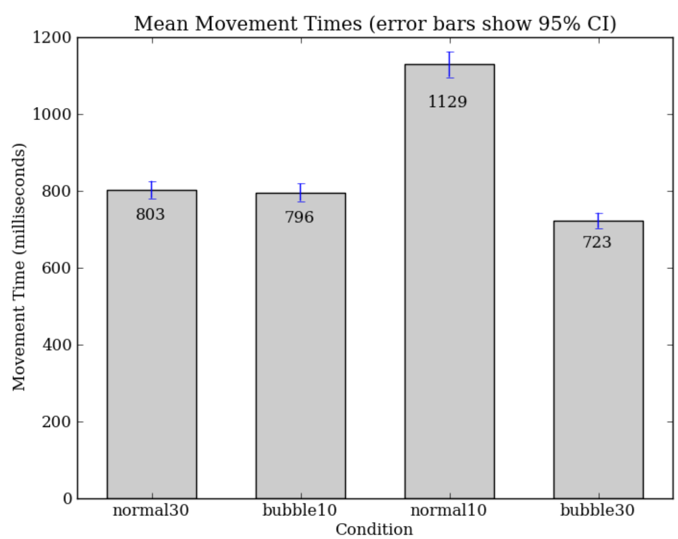
Calculate with the help of the standard error SE.

Standard Deviation: measures variability of individual data points.

Standard Error: measures variability of means

$$SE = \frac{SD}{\sqrt{N}}$$

$$95\%CI = M \pm 1.96 \times SE$$



Draw Conclusions

What is the scope of the finding?

Are there other parameters at play?

Internal validity

Does the experiment reflect real use?

External 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

Midterm Review

Midterm on 3/14

In class. 75 minutes

Closed book & notes

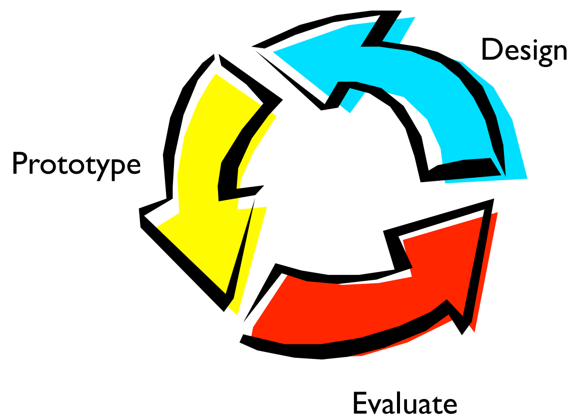
Test is long - so be strategic

- 1st pass: Read through entire test, give immediate answers
- 2nd pass: Go back, answer questions requiring more time

Extra Office Hours

- Tue 3/13 11am-noon, 3-4pm (405 Soda Hall)
- Tue 3/13 2-3pm (BiD)

Interface Design Cycle



Task Analysis & Contextual Inquiry

Observe existing practices

Create scenarios of actual use

Create models to gain insight into work processes

CS247, Stanford, 2006

http://www-personal.umich.edu/~chrisli/m2.html

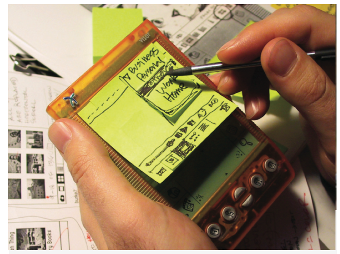
Rapid Prototyping

Build a mock-up of design
(or more!)

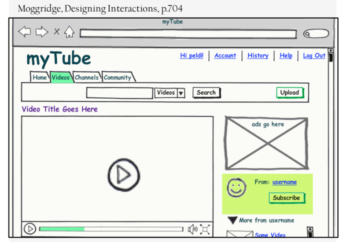
Low fidelity techniques
Paper sketches
Cut, copy, paste
Video segments

Interactive prototyping tools
HTML, Flash, Javascript,
Visual Basic, C#, etc.

UI builders
Interface Builder, Visual Studio, NetBeans



Moggridge, Designing Interactions, p704



http://www.balsamiq.com/products/mockups/examples#wiki

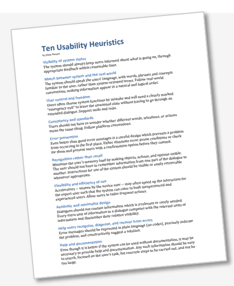
Evaluation

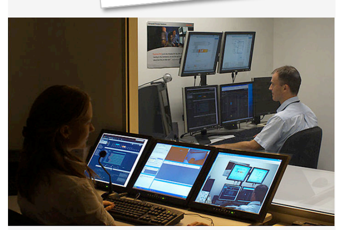
Evaluate analytically (no users)

Test with real target users

Low-cost techniques
expert evaluation
walkthroughs

Higher cost
Controlled usability study





http://www.laurasmith.info/UsabilityTest.jpg

Comparison

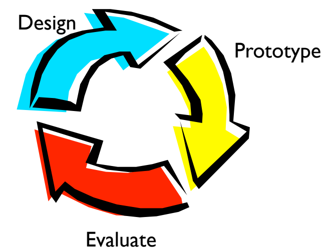
Focus differs

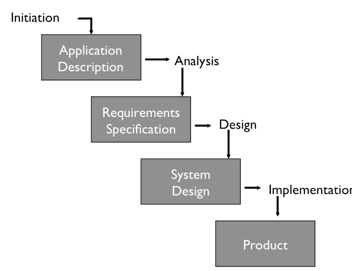
WF has no feedback

High cost of fixing errors: increases by 10x at each stage

Iterative design finds problems earlier

True for modern web applications?





IDEO's Brainstorming Rules

1. Sharpen the Focus
2. Playful Rules
3. Number your Ideas
4. Build and Jump
5. The Space Remembers
6. Stretch Your Mental Muscles
7. Get Physical

Aim for quantity

Hope for quality

Task Analysis Questions

1. Who is going to use system?
2. What tasks do they now perform?
3. What tasks are desired?
4. How are the tasks learned?
5. Where are the tasks performed?
6. What's the relationship between user & data?
7. What other tools does the user have?
8. How do users communicate with each other?
9. How often are the tasks performed?
10. What are the time constraints on the tasks?
11. What happens when things go wrong?

Goals of Contextual Inquiry

Method:

"Go where the customer works, observe the customer as she works, and talk to the customer about their work" [Holtzblatt]

Goals:

Get inside the user's head

See their tasks the way they do

A middle ground between pure observation and pure interview

Guideline: Master-Apprentice Model

Allows user to teach us what they do
- Skill knowledge is usually tacit (can't put it in books)
- Sometimes literal apprenticeship is best


→


Matsushita Home Bakery – First automatic bread maker to have twist/stretch motion [Nonaka 95]

Principles of Contextual Inquiry

1. Context
2. Partnership
3. Interpretation
4. Focus

Personas (from Cooper)

"Hypothetical Archetypes"

Archetype: (American Heritage)

An original model or type after which other similar things are patterned; a prototype

An ideal example of a type; quintessence

A precise description of user in terms

Capabilities, inclinations, background

Goals (not tasks)

Why Personas?

It's hard to reason about users in aggregate, and impossible to please everyone.

General users have too many conflicting goals.

http://simpsons.wikia.com/wiki/File:TheHomer.png

"… the term **affordance** refers to the *perceived* and *actual* properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used.

Some affordances obvious

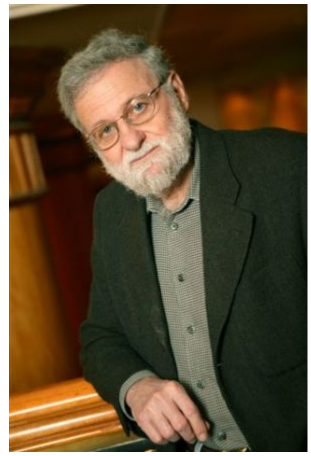
Knobs afford turning

Buttons afford pushing

Glass can be seen through

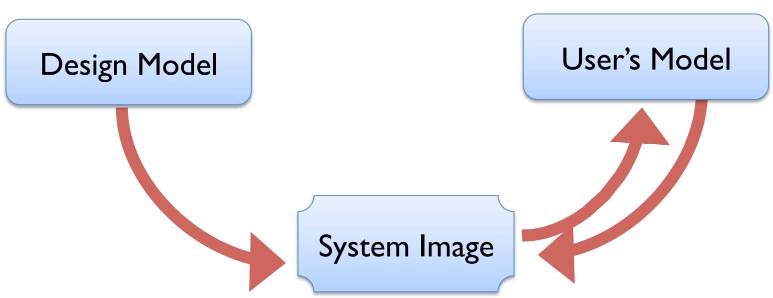
Some affordances learned

Glass breaks easily



The Design of Everyday Things.
Don Norman

Review Conceptual Models



Designers model may not match user's model

Users get model from experience & usage

Users only work with system image, not with designer

What if the two models don't match?

1. Make Controls Visible

2. Make Sure Mapping is Clear

Mapping: Relationship between controls and their result

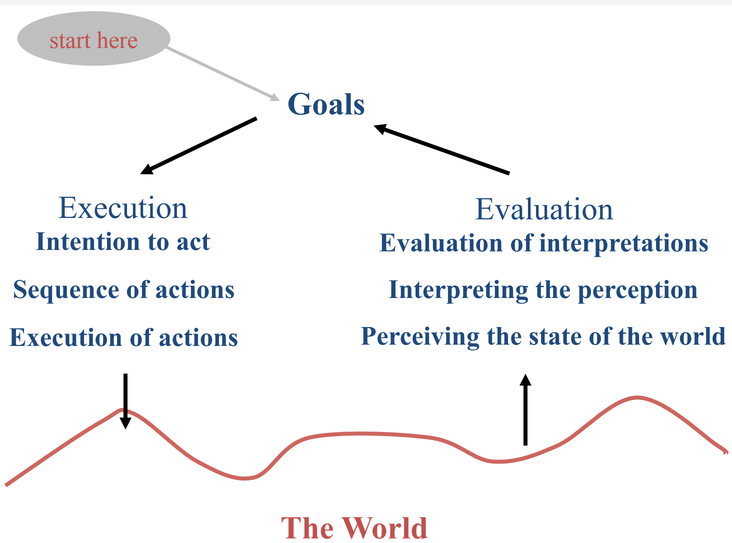
Mercedes S500 Car Seat Controller

3. Provide Feedback

People press >> 1 time

Unclear if system has registered the button press

Action Cycle



Direct Manipulation

An interface that behaves as though the interaction was with a real-world object rather than with an abstract system

Central ideas

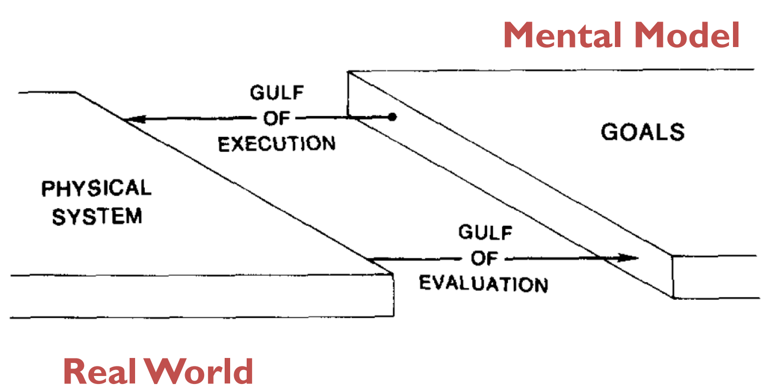
Visibility of the objects of interest

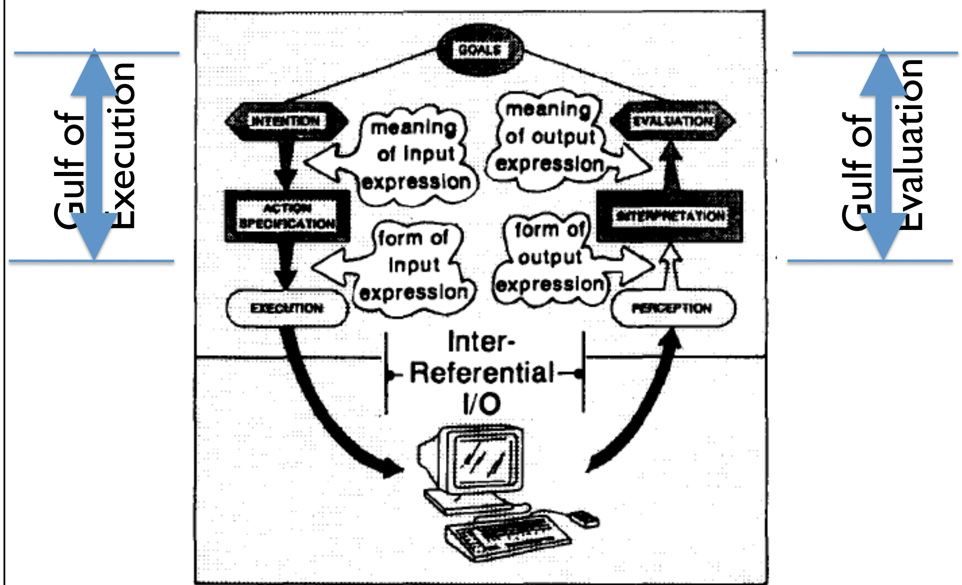
Rapid, reversible, incremental actions

Manipulation by pointing and moving

Immediate and continuous display of results

Gulfs of Execution & Evaluation

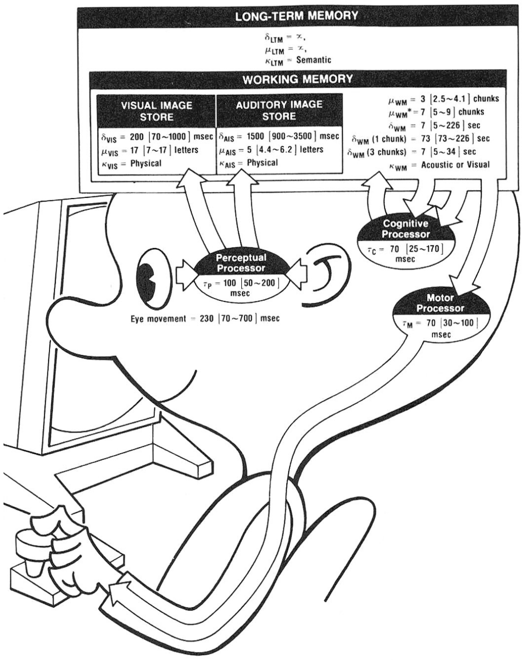




Modes: Definition

The same user actions have different effects in different situations.

Examples?



Human Info. Processor

Processors:

- Perceptual
- Cognitive
- Motor

Memory:

- Working memory
- Long-term memory

Unified model

Probably inaccurate

Predicts perf. well

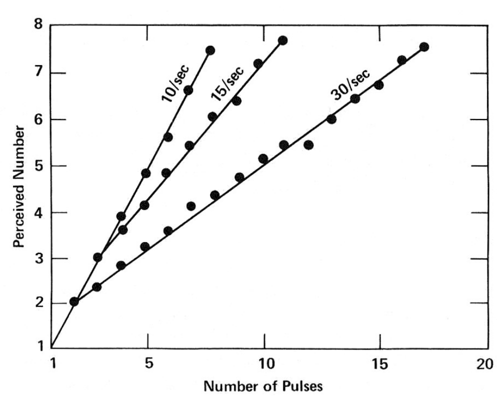
Very influential

Perceptual Processor

Cycle time

Quantum experience: 100ms

Percept fusion



Working Memory

Access in chunks

Task dependent construct

7 +/- 2 (Miller)

Decay

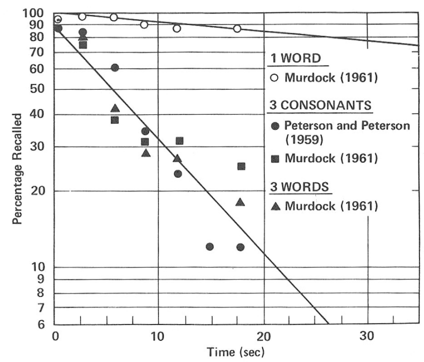
Content dependant

1 chunk 73 sec

3 chunks 7 sec

Attention span

Interruptions > decay time



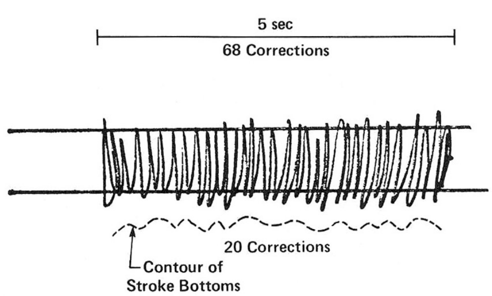
Motor Processor

Receive input from the cognitive processor

Execute motor programs

Pianist: up to 16 finger movements per second

Point of no-return for muscle action

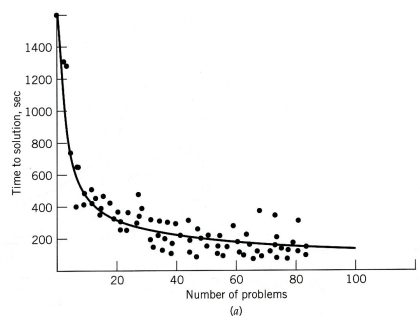


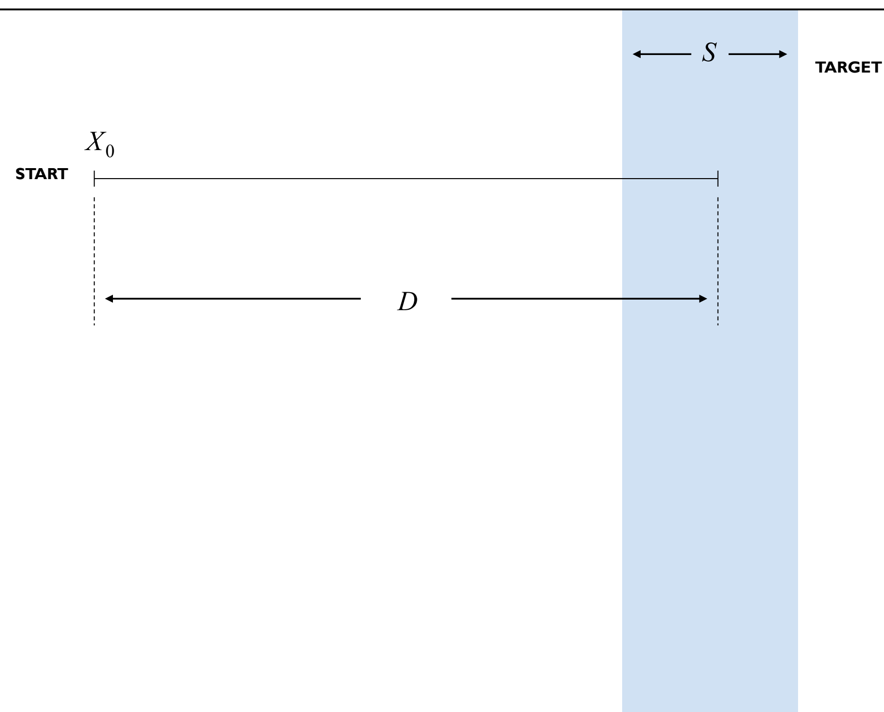
Power Law of Practice

Task time on the nth trial follows a power law

$$T_n = T_1 n^{-a} + c$$

You get faster the more times you do something!





Fitts' Law

$$T = a + b\log_2(D/S + 1)$$

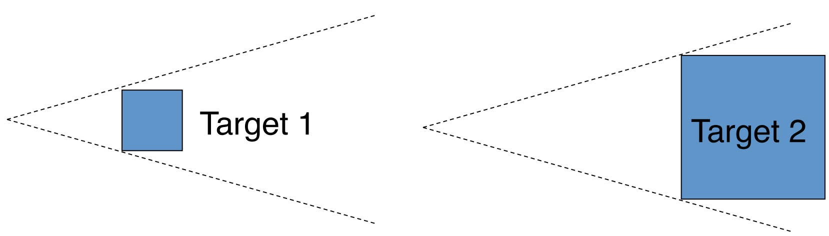
a, b = constants (empirically derived)
D = distance
S = size

ID is Index of Difficulty = $\log_2(D/S+1)$

- Models well-rehearsed selection task
- T increases as the ***distance*** to the target increases
- T decreases as the ***size*** of the target increases

Considers Distance and Target Size

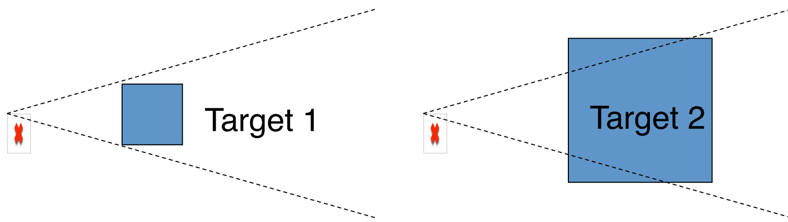
$$T = a + b\log_2(D/S + 1)$$



Same ID → Same Difficulty

Considers Distance and Target Size

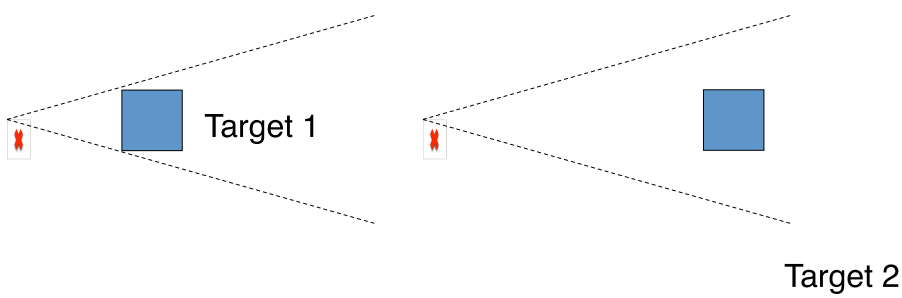
$$T = a + b\log_2(D/S + 1)$$



Smaller ID → Easier

Considers Distance and Target Size

$$T = a + b\log_2(D/S + 1)$$



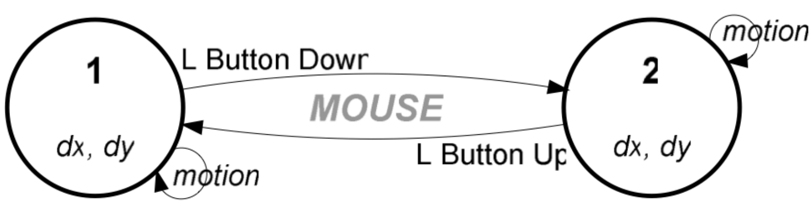
Larger ID → Harder

3-State Model of Input (Buxton)

State	Description
0	*Out Of Range:* The device is not in its physical tracking range.
1	*Tracking:* Device motion moves only the cursor.
2	*Dragging:* Device motion moves objects on the screen.

(Table from Hinckley Reading)

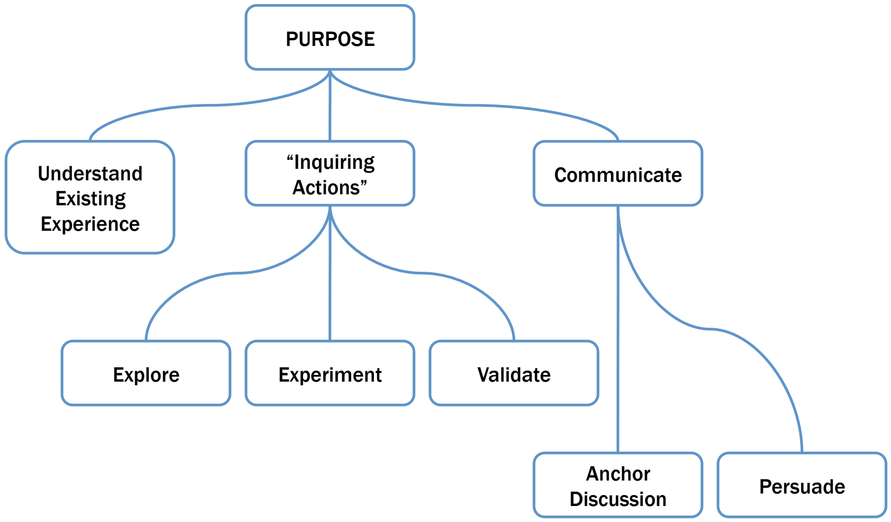
Mouse



(Figure from Hinckley Reading)

State	Description
0	*Out Of Range:* The device is not in its physical tracking range.
1	*Tracking:* Device motion moves only the cursor.
2	*Dragging:* Device motion moves objects on the screen.

Prototyping



Fidelity in Prototyping

Fidelity refers to the level of detail

High fidelity

Prototypes look like the final product

Low fidelity

Artists renditions with many details missing

Paper Prototypes are low-fidelity.
What about software?

Hi-Fi Disadvantages

Distort perceptions of the tester

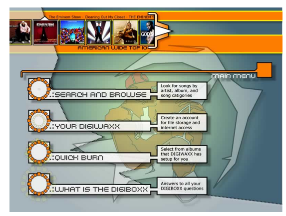
Formal representation indicates "finished" nature

People comment on color, fonts, and alignment

Discourages major changes

Testers don't want to change a "finished" design

Sunk-cost reasoning: Designers don't want to lose effort put into creating hi-fi design



Engineering Interfaces

User Interface Components

Each component is an object with

Bounding box

Paint method for drawing itself

Drawn in the component's coordinate system

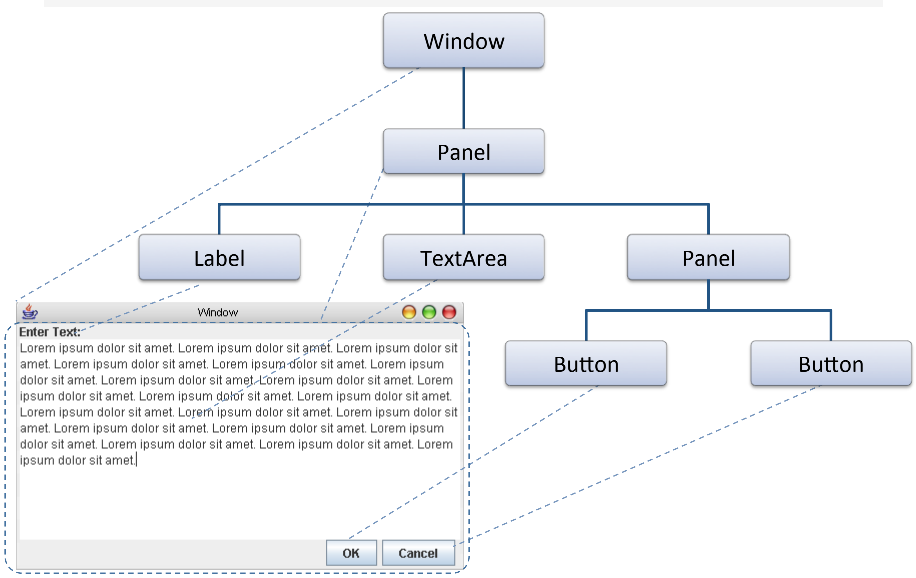
Callbacks to process input events

Mouse clicks, typed keys

Java:
```
public void paint(Graphics g) {
   g.fillRect(…); // interior
   g.drawString(…); // label
   g.drawRect(…); // outline
}
```

Cocoa:
(void)**drawRect:**(NSRect)*rect*

Layout: Containment Hierarchy



Anatomy of an Event

Encapsulates info needed for handlers to react to input

Event Type (mouse moved, key down, etc)

Event Source (the input component)

Timestamp (when did event occur)

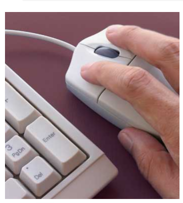
Modifiers (Ctrl, Shift, Alt, etc)

Event Content

Mouse: x,y coordinates, button pressed, # clicks

Keyboard: which key was pressed

Event Dispatch Loop



Mouse moved (t_0,x,y)

Event Queue
- Queue of input events

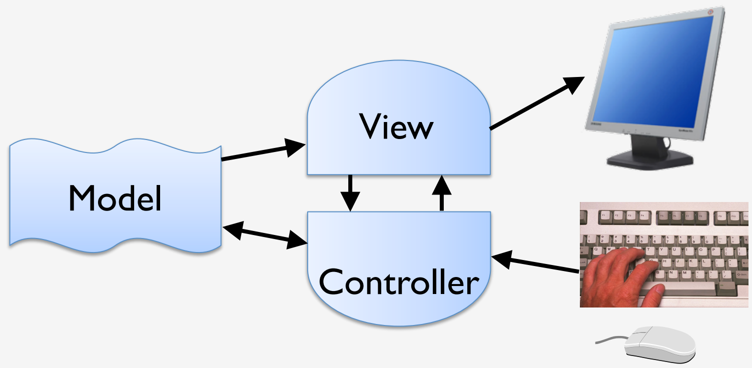
Event Loop (runs in dedicated thread)
- Remove next event from queue
- Determine event type
- Find proper component(s)
- Invoke callbacks on components
- Repeat, or wait until event arrives

Component
- Invoked callback method
- Update application state
- Request repaint, if needed

Model-View-Controller

OO Architecture for interactive applications

introduced by Smalltalk developers at PARC ca. 1983



Why MVC?

Combining MVC into one class will not scale

model may have more than one view

each is different and needs update when model changes

Separation eases maintenance and extensibility

easy to add a new view later

model info can be extended, but old views still work

can change a view later, e.g., draw shapes in 3D

flexibility of changing input handling when using separate controllers

Changing the Display

Erase and redraw

using background color to erase fails

drawing shape in new position loses ordering

Damage / Redraw Method

View informs windowing system of areas that are damaged

does not redraw them right away…

Windowing system

batches updates

clips them to visible portions of window

Next time waiting for input

windowing system calls Repaint() method

passes region that needs to be updated

What is a thread?

A **thread** is a **partial virtual machine**.

Each thread has its own stack (and local variables) but shares its heap with other threads in the same application.

Threads can be independently scheduled by the OS/VM.

```
for (i=0; i<n; i++)
{
  tmp = A[i];
  A[i] = B[i];
  B[i] = tmp;
}
```

Thread1 Thread2

Why use multithreading for UIs?

Not all code can complete quickly inside an event handler. Examples?

Updating the UI from another thread

All common UI frameworks have a single UI thread

You are only allowed to modify the UI from the main thread.

Two fundamental rules:

Do not block the UI thread

Background threads they **must not modify the UI.**

Solution: When worker thread completes, request update back in the UI thread.

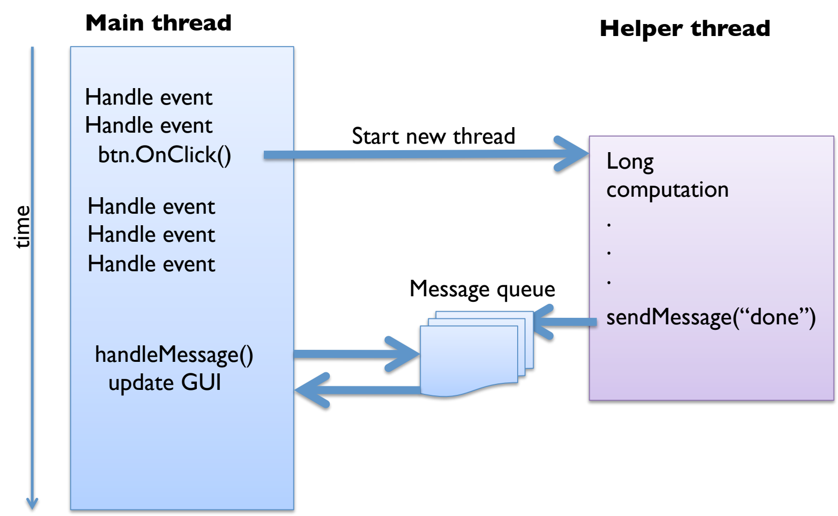
How to properly update the UI

Almost all GUI frameworks offer some convenient mechanism to notify the main thread from another thread.

Android has at least three such mechanisms:

1. Call View.post(Runnable) from worker thread
2. Subclass AsyncTask – creates threads behind the scenes
3. Send messages in one thread with Handler. sendMessage() – message is received in another thread (like IPC)

Handler.sendMessage Example



Usability Testing Methods

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

Usability Heuristics

"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

Nielsen's Ten 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, recover from errors
H2-10: Help and documentation

Phases of Heuristic Eval. (1-2)

1) Pre-evaluation training
Provide the evaluator with domain knowledge if needed

2) Evaluation
Individuals evaluate interface then aggregate results
Compare interface elements with heuristics

Work in 2 passes
First pass: get a feel for flow and scope
Second pass: focus on specific elements

Each evaluator produces list of problems
Explain why with reference to heuristic or other information
Be specific and list each problem separately

Phases of Heuristic Eval. (3-4)

3) Severity rating

Establishes a ranking between problems

Cosmetic, minor, major and catastrophic

First rate individually, then as a group

4) Debriefing

Discuss outcome with design team

Suggest potential solutions

Assess how hard things are to fix

Number of Evaluators

Single evaluator achieves poor results

Only finds 35% of usability problems

5 evaluators find ~ 75% of usability problems

Why not more evaluators???? 10? 20?

Adding evaluators costs more

Many evaluators won't find many more problems

But always depends on market for product:

popular products → high support cost for small bugs

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 IV could have caused change in DVs

Confounds make it difficult/impossible to draw conclusions

Random variables

Attributes that are randomly sampled

Increases generalizability

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.

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

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

HE vs. User Testing

HE is much faster
1-2 hours each evaluator vs. days-weeks

HE doesn't require interpreting user's actions

User testing is far more accurate
Takes into account actual users and tasks
HE may miss problems & find "false positives"

Good to alternate between HE & user-based testing
Find different problems
Don't waste participants

The Three Belmont Principles

Respect for Persons
Have a meaningful consent process: give information, and let prospective subjects freely chose to participate

Beneficience
Minimize the risk of harm to subjects, maximize potential benefits

Justice
Use fair procedures to select subjects
(balance burdens & benefits)

To ensure adherence to principles, most schools require Institutional Review Board approval of research involving human subjects.

Descriptive Statistics

Continuous data:

Central tendency

mean,median,mode

Dispersion

Range (max-min)

Standard deviation

Shape of distribution

Skew, Kurtosis

Categorical data:

Frequency distributions

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Mean

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

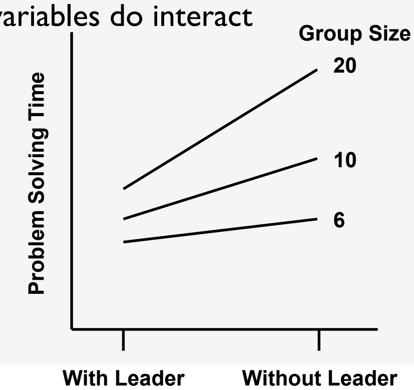
Standard Deviation

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]

Are the Results Meaningful?

Hypothesis testing

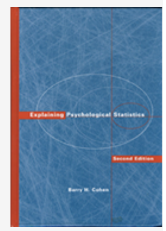
Hypothesis: Manipulation of IV effects DV in some way
Null hypothesis: Manipulation of IV has no effect on DV
Null hypothesis assumed true unless statistics allow us to reject it

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

What does p>0.05 mean?

No statistically significant difference (at 5% level)

Are the two conditions thus equivalent?

NO! We DID observe differences.

But can't be sure they are not due to chance.

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