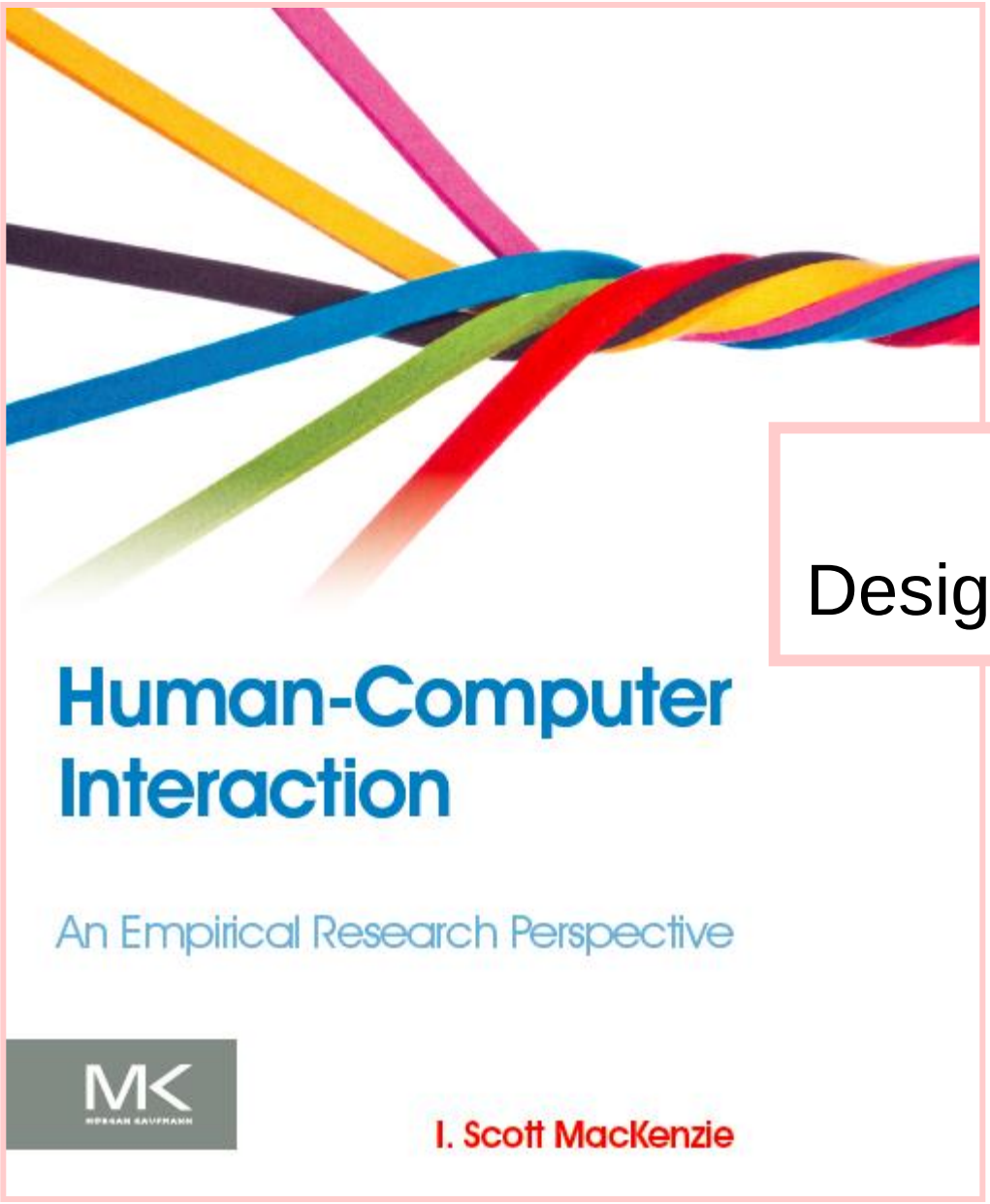


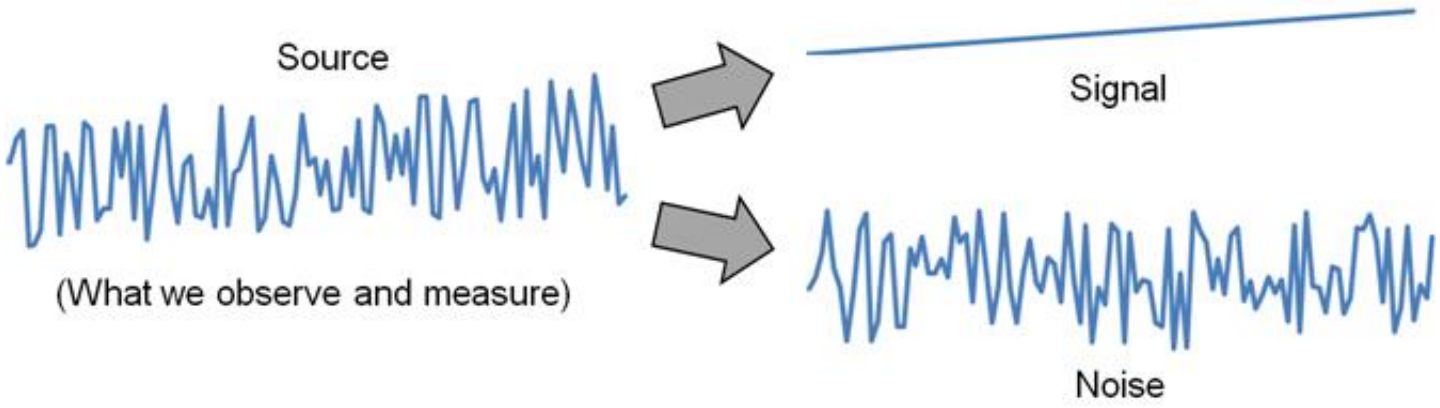
Chapter 5
Designing HCI Experiments

Introduction

- Learning to conduct and design an experiment is a skill required of all researchers in HCI
- *Experiment design* is the process of deciding what variables to use, what tasks and procedures to use, how many participants to use and how to solicit them, and so on

Signal and Noise Metaphor

- Signal and noise metaphor for experiment design:



- Signal → a variable of interest
- Noise → everything else (random influences)
- Experiment design seeks to enhance the signal, while minimizing the noise

Methodology

- *Methodology* is the way an experiment is designed and carried out
- Methodology is critical:

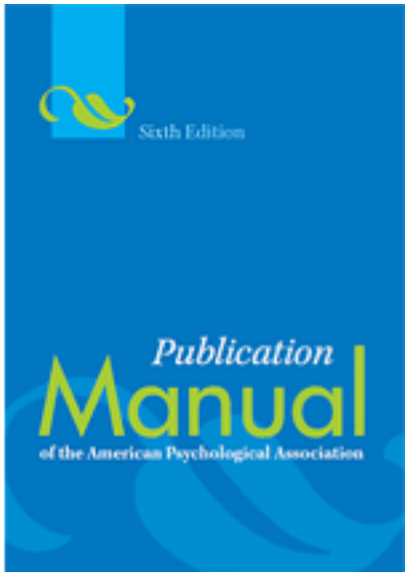
> Science is method. Everything else is commentary.[1]

- What methodology?
- Don't just make it up because it seems reasonable
- Follow standards for experiments with human participants (next slide)

[1] This quote from Allen Newell was cited and elaborated on by Stuart Card in an invited talk at the ACM's SIGCHI conference in Austin, Texas (May 10, 2012).

APA

- American Psychological Society (APA) is the pre-dominant organization promoting research in psychology – the improvement of research methods and conditions and the application of research findings (http://www.apa.org/)
- *Publication Manual of the APA*[1], first published in 1929, teaches about the writing process and, implicitly, about the methodology for experiments with human participants
- Recommended by major journals in HCI



[1] APA. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

Ethics Approval

- *Ethics approval* is a crucial step that precedes every HCI experiment
- HCI experiments involve humans, thus…

> Researchers must respect the safety, welfare, and dignity of human participants in their research and treat them equally and fairly.[1]

- Proposal submitted to ethics review committee
- Criteria for approval:
- research methodology
  - risks or benefits
  - the right not to participate, to terminate participation, etc.
  - the right to anonymity and confidentiality

[1] http://www.yorku.ca/research/students/index.html

Getting Started With Experiment Design

- Transitioning from the creative work in formulating and prototyping ideas to experimental research is a challenge
- Begin with…

> What are the experimental variables?

- Remember research questions:

> Can a task be performed more quickly with my new interface than with an existing interface?

- Properly formed research questions inherently identify experimental variables (can you spot the independent variable and the dependent variable in the question above?)

Independent Variable

- An *independent variable* (IV) is a circumstance or characteristic that is manipulated in an experiment to elicit a change in a human response while interacting with a computer.
- "Independent" because it is independent of participant behavior (i.e., there is nothing a participant can do to influence an independent variable)
- Examples:
  - interface, device, feedback mode, button layout, visual layout, age, gender, background noise, expertise, etc.
- The terms *independent variable* and *factor* are synonymous

Test Conditions

- An independent variable (IV) must have at least two levels
- The levels, values, or settings for an IV are the *test conditions*
- Name both the factor (IV) and its levels (test conditions):

Factor (IV)	Levels (test conditions)
Device	mouse, trackball, joystick
Feedback mode	audio, tactile, none
Task	pointing, dragging
Visualization	2D, 3D, animated
Search interface	Google, custom

Human Characteristics

- Human characteristics are *naturally occurring attributes*
- Examples:
  - Gender, age, height, weight, handedness, grip strength, finger width, visual acuity, personality trait, political viewpoint, first language, shoe size, etc.
- They are legitimate independent variables, but they cannot be “manipulated” in the usual sense
- Causal relationships are difficult to obtain due to unavoidable confounding variables

How Many IVs?

- An experiment must have at least one independent variable
- Possible to have 2, 3, or more IVs
- But the number of "effects" increases rapidly with the size of the experiment:

Independent Variables	Effects					Total
	Main	2-way	3-way	4-way	5-way	
1	1	-	-	-	-	1
2	2	1	-	-	-	3
3	3	3	1	-	-	7
4	4	6	3	1	-	14
5	5	10	6	3	1	25

- Advice: Keep it simple (1 or 2 IVs, 3 at the most)

Dependent Variable

- A *dependent variable* is a measured human behaviour (related to an aspect of the interaction involving an independent variable)
- "Dependent" because it depends on what the participant does
- Examples:
  - task completion time, speed, accuracy, error rate, throughput, target re-entries, task retries, presses of backspace, etc.
- Dependent variables must be clearly defined
  - Research must be reproducible!

Unique DVs

- Any observable, measurable behaviour is a legitimate dependent variable (provided it has the potential to reveal differences among the test conditions)
- So, feel free to "roll your own"
- Example: *negative facial expressions*[1]
  - Application: user difficulty with mobile games
  - Events logged included frowns, head shaking
  - Counts used in ANOVA, etc.
  - Clearly defined → reproducible

[1] Duh, H. B.-L., Chen, V. H. H., & Tan, C. B. (2008). Playing different games on different phones: An empirical study on mobile gaming. *Proceedings of MobileHCI 2008*, 391-394, New York: ACM.

Data Collection

- Obviously, the data for dependent variables must be collected in some manner
- Ideally, engage the experiment software to log timestamps, key presses, button clicks, etc.
- Planning and pilot testing important
- Ensure conditions are identified, either in the filenames or in the data columns
- Examples: (next two slides)

TextInputHuffman-P01-D99-B06-S01.sd2

```
min_keystrokes,keystrokes,presented_characters,transcribed_characters, ...
55, 59, 23, 23, 29.45, 0, 9.37, 0.0, 2.5652173913043477, 93.22033898305085
61, 65, 26, 26, 30.28, 0, 10.3, 0.0, 2.5, 93.84615384615384
85, 85, 33, 33, 48.59, 0, 8.15, 0.0, 2.5757575757575757, 100.0
67, 71, 28, 28, 33.92, 0, 9.91, 0.0, 2.5357142857142856, 94.36619718309859
61, 70, 24, 24, 39.44, 0, 7.3, 0.0, 2.9166666666666665, 87.14285714285714
```

Control Variable

- A *control variable* is a circumstance (not under investigation) that is kept constant while testing the effect of an independent variable
- More control means the experiment is less generalizable (i.e., less applicable to other people and other situations)
- Research question: Is there an effect of font color or background color on reading comprehension?
  - Independent variables: font color, background color
  - Dependent variable: comprehension test scores
  - Control variables
    - Font size (e.g., 12 point)
    - Font family (e.g., Times)
    - Ambient lighting (e.g., fluorescent, fixed intensity)
    - Etc.

Random Variable

- A *random variable* is a circumstance that is allowed to vary randomly
- More variability is introduced in the measures (that's bad!), but the results are more generalizable (that's good!)
- Research question: Does user stance affect performance while playing *Guitar Hero?*
  - Independent variable: stance (standing, sitting)
  - Dependent variable: score on songs
  - Random variables
    - Prior experience playing a real musical instrument
    - Prior experience playing *Guitar Hero*
    - Amount of coffee consumed prior to testing
    - Etc.

Control vs. Random Variables

- There is a trade-off which can be examined in terms of internal validity and external validity (see below)

Variable	Advantage	Disadvantage
Random		
Control		

Confounding Variable

- A *confounding variable* is a circumstance that varies systematically with an independent variable
- Should be considered, else the results are misleading
- Research question: In an eye tracking application, is there an effect of "camera distance" on task completion time?
  - Independent variable: Camera distance (near, far)
    - Near camera (A): inexpensive camera mounted on eye glasses
    - Far camera (B): expensive camera mounted above system display
  - Dependent variable: task completion time
  - But, "camera" is a confounding variable: camera A for the near setup, camera B for the far setup
  - Are the effects due to camera distance or to some aspect of the different setups?

Experiment Task

- Recall the definition of an independent variable:
  - a circumstance or characteristic that is manipulated in an experiment to *elicit a change in a human response* while interacting with a computer
- The experiment task must "elicit a change"
- Qualities of a good task: *represent*, *discriminate*
  - Represent activities people do with the interface
    - Improves external validity (but may compromise internal validity)
  - Discriminate among the test conditions
    - Increases likelihood of a statistically significant outcome (i.e., the sought-after "change" occurs)

Task Examples

- Usually the task is self-evident (follows directly from the research idea)
- Research idea → a new graphical method for entering equations in a spreadsheet
  - Experiment task → insert an equation using (a) the graphical method and (b) the conventional method
- Research idea → an auditory feedback technique for programming a GPS device
  - Experiment task → program a destination location into a GPS device using (a) the auditory feedback method and (b) the conventional method

Knowledge-based Tasks

- Most experiment tasks are *performance-based* or *skill-based* (e.g., inserting an equation, programming a destination location; see previous slide)
- Sometimes the task is *knowledge-based* (e.g., "Use an Internet search interface to find the birth date of Albert Einstein.")
- In this case, participants become contaminated (in a sense) after the first run of task, since they have acquired the knowledge
- Experimentally, this poses problems (beware!)
- A creative approach is needed (e.g., for the other test condition, slightly change the task; "…of William Shakespeare")

Procedure

- The *procedure* encompasses everything that occurs with participants
- The procedure includes the experiment task (obviously), but everything else as well…
  - Arriving, welcoming
  - Signing a consent form
  - Instructions given to participants about the experiment task (next slide)
  - Demonstration trials, practice trials
  - Rest breaks
  - Administering of a questionnaire or an interview

Instructions

- Very important (best to prepare in advance; write out)
- Often the goal in the experiment task is “to proceed as quickly and accurately as possible but at a pace that is comfortable”
- Other instructions are fine, as per the goal of the experiment or the nature of the tasks, but…
- Give the same instructions to all participants
- If a participant asks for clarification, do not change the instructions in a way that may cause the participant to behave differently from the other participants

Participants

- Researchers want experimental results to apply to people not actually tested – a population
- Population examples:
  - Computer-literate adults, teenagers, children, people with certain disabilities, left-handed people, engineers, musicians, etc.
- For results to apply generally to a population, the participants used in the experiment must be…
  - Members of the desired population
  - Selected at random from the population
- True random sampling is rarely done (consider the number and location of people in the population examples above)
- Some form of *convenience sampling* is typical

How Many Participants?

- Too few → experimental effects fail to achieve statistical significance
- Too many → statistical significance for effects of no practical value
- The correct number… (drum roll please)
  - Use the same number of participants as used in similar research[1]

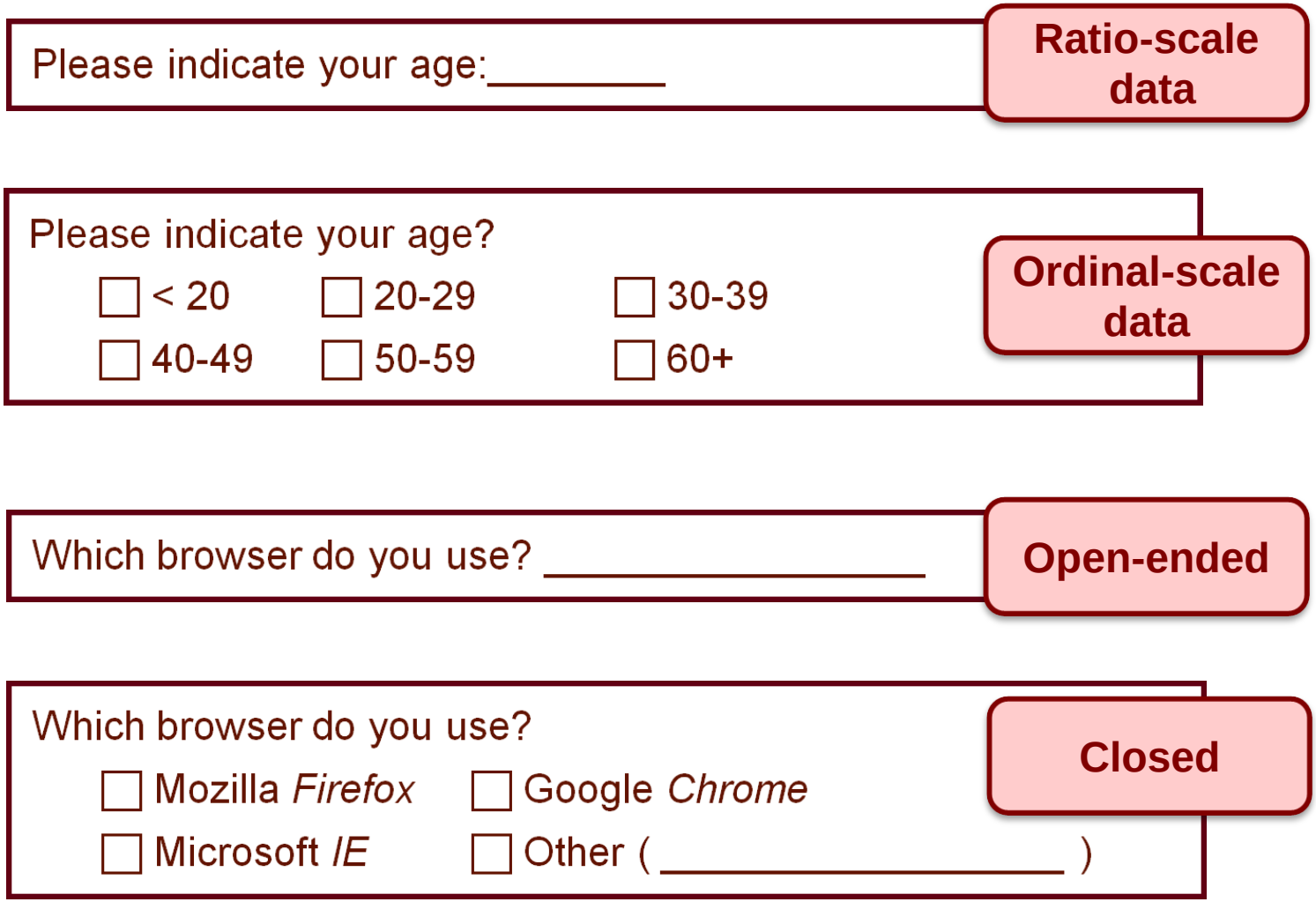
[1] Martin, D. W. (2004). *Doing psychology experiments* (6th ed.). Pacific Grove, CA. Belmont, CA: Wadsworth.

Questionnaires

- Questionnaires are used in most HCI experiments
- Two purposes
  - Collect information about the participants
    - Demographics (gender, age, first language, handedness, visual acuity, etc.)
    - Prior experience with interfaces or interaction techniques related to the research
  - Solicit feedback, comments, impressions, suggestions, etc., about participants' use of the experimental apparatus
- Questionnaires, as an adjunct to experimental research, are usually brief

Information Questions

- Questions constructed according to how the information will be used



Participant Feedback

- Using NASA Task Load Index (TLX):

Frustration: I felt a high level of insecurity, discouragement, irritation, stress, or annoyance.

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly agree

- ISO 9241-9:

Eye fatigue:

1	2	3	4	5	6	7
Very high						Very low

Within-subjects, Between-subjects

- Two ways to assign conditions to participants:
  - *Within-subjects* → each participant is tested on each condition
  - *Between-subjects* → each participant is tested on one condition only
  - Example: An IV with three test conditions (A, B, C):

Within-subjects

Participant	Test Condition		
1	A	B	C
2	A	B	C

Between-subjects

Participant	Test Condition
1	A
2	A
3	B
4	B
5	C
6	C

Within-subjects, Between-subjects (2)

- Within-subjects advantages
  - Fewer participants (easier to recruit, schedule, etc.)
  - Less “variation due to participants”
  - No need to balance groups (because there is only one group!)
- Within-subjects disadvantage
  - Order effects (i.e., interference between conditions)
- Between-subjects advantage
  - No order effects (i.e., no interference between conditions)
- Between-subjects disadvantage
  - More participants (harder to recruit, schedule, etc.)
  - More “variation due to participants”
  - Need to balance groups (to ensure they are more or less the same)

Within-subjects, Between-subjects (3)

- Sometimes…
  - A factor must be assigned within-subjects
    - Examples: Block, session (if learning is the IV)
  - A factor must be assigned between-subjects
    - Examples: gender, handedness
  - There is a choice
    - In this case, the balance tips to within-subjects (see previous slide)
- With two factors, there are three possibilities:
    - both factors within-subjects
    - both factors between-subjects
    - one factor within-subjects + one factor between-subjects (this is a *mixed design*)

Order Effects, Counterbalancing

- Only relevant for within-subjects factors
- The issue: *order effects* (aka *learning effects*, *practice effects*, *fatigue effects*, *sequence effects*)
- Order effects offset by *counterbalancing*:
  - Participants divided into groups
  - Test conditions are administered in a different order to each group
  - Order of administering test conditions uses a Latin square
  - Distinguishing property of a Latin square → each condition occurs precisely once in each row and column (next slide)

Latin Squares

2 x 2

A	B
B	A

3 x 3

A	B	C
B	C	A
C	A	B

4 x 4

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

5 x 5

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

Balanced Latin Square

- With a balanced Latin square, each condition precedes and follows each other condition an equal number of times
- Only possible for even-orders
- Top row pattern: A, B, *n*, C, *n* – 1, D, *n* – 2, …

4 x 4

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

6 x 6

A	B	F	C	E	D
B	C	A	D	F	E
C	D	B	E	A	F
D	E	C	F	B	A
E	F	D	A	C	B
F	A	E	B	D	C

Example

- An experimenter seeks to determine if three editing methods (A, B, C) differ in the amount of time to do a common editing task:

> Replace one 5-letter word with another, starting one line away.

- Conditions are assigned within-subjects
- Twelve participants are recruited and divided into three groups (4 participants/group)
- Methods administered using a 3×3 Latin Square (2 slides back)
- Results (next slide)

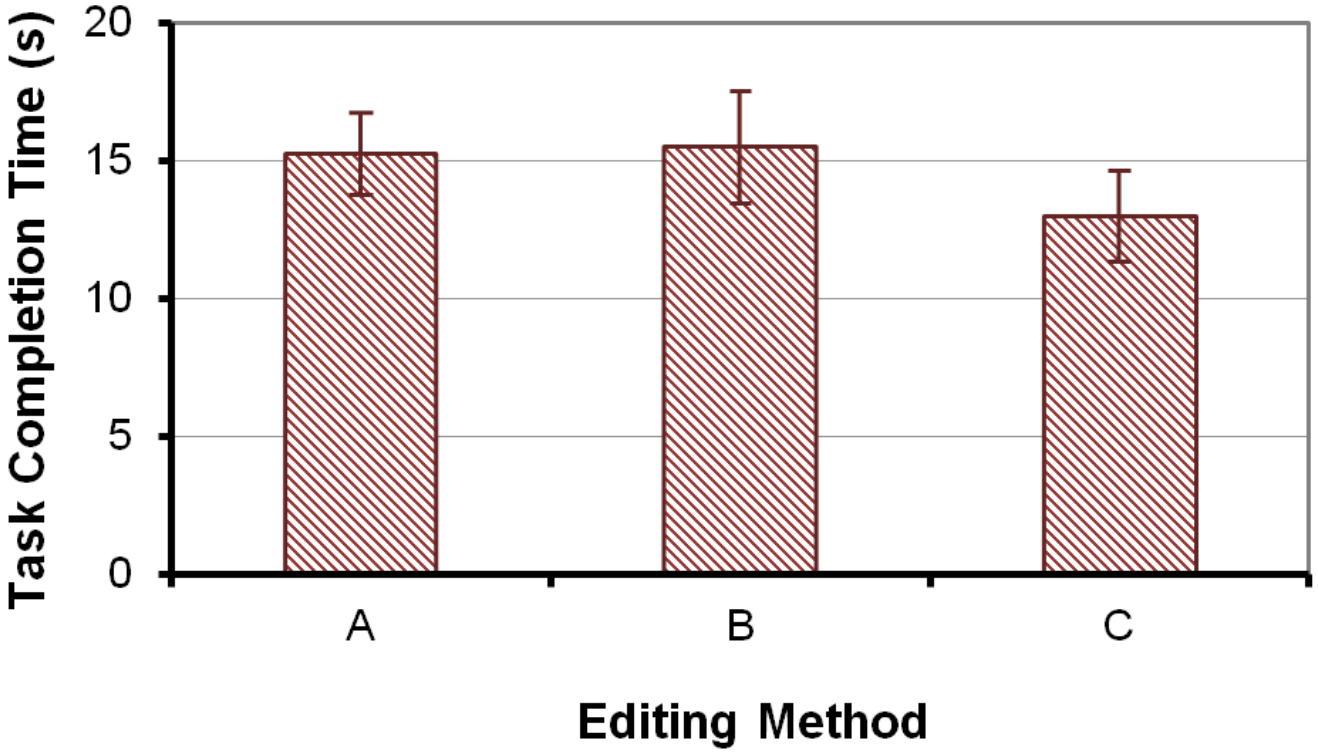
Results - Data

Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

Group effect is small

∴ Counterbalancing worked!

Results - Chart

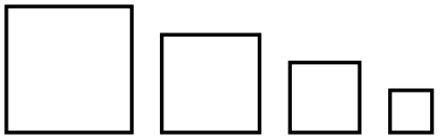


Other Techniques

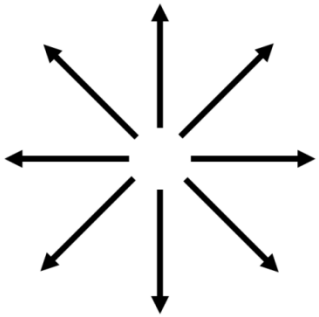
- Instead of using a Latin square, all orders (*n*!) can be used; 3 × 3 case →
- Conditions can be randomized
- Randomizing best if the tasks are brief and repeated often; examples (see below)

A	B	C
A	C	B
B	C	A
B	A	C
C	A	B
C	B	A

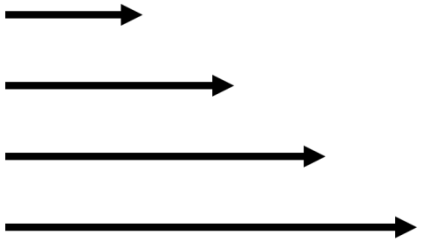
Target size



Movement direction



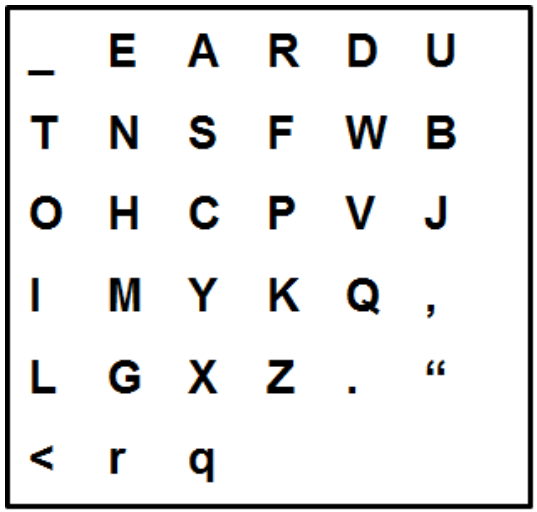
Movement distance



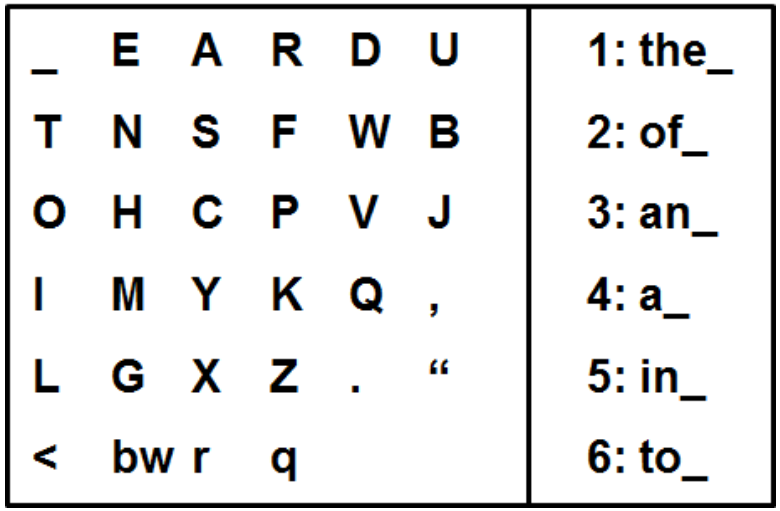
Asymmetric Skill Transfer

- A phenomenon known as *asymmetric skill transfer* sometimes occurs in within-subjects designs
- Figures in next three slides demonstrate
- They are presented in the slides without an explanation (to be discussed in class)
- Please see **HCI:ERP** for complete details and discussion (including how to avoid asymmetric skill transfer)

LO Keyboard

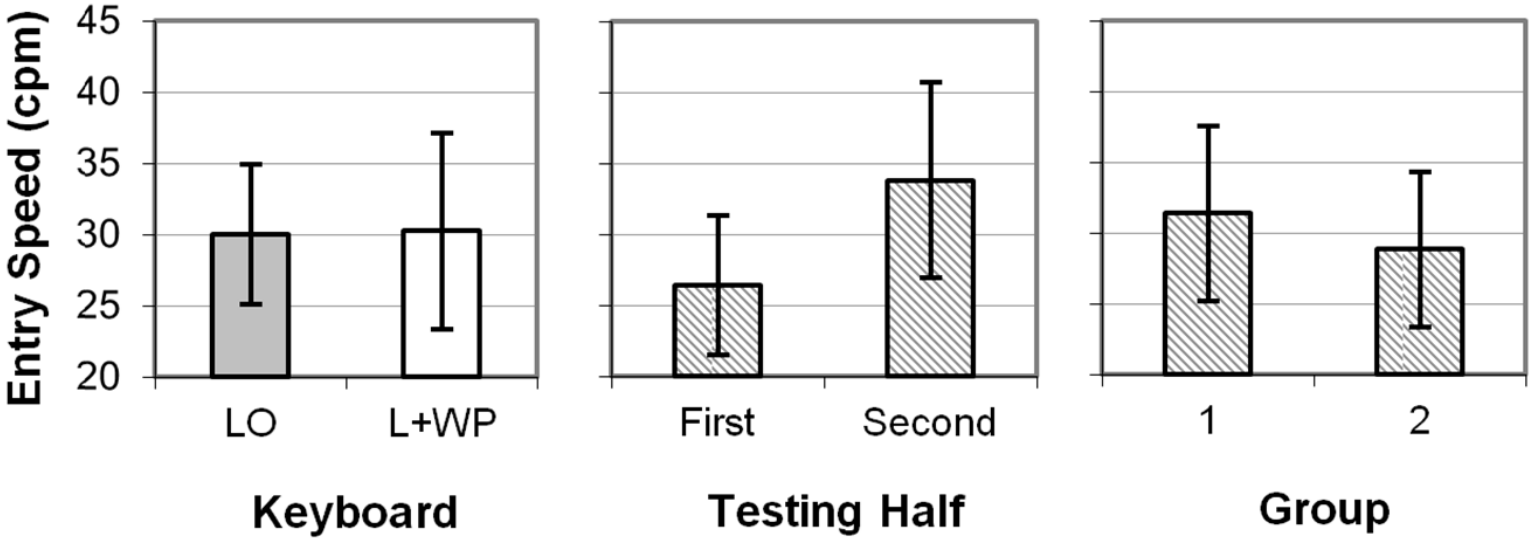


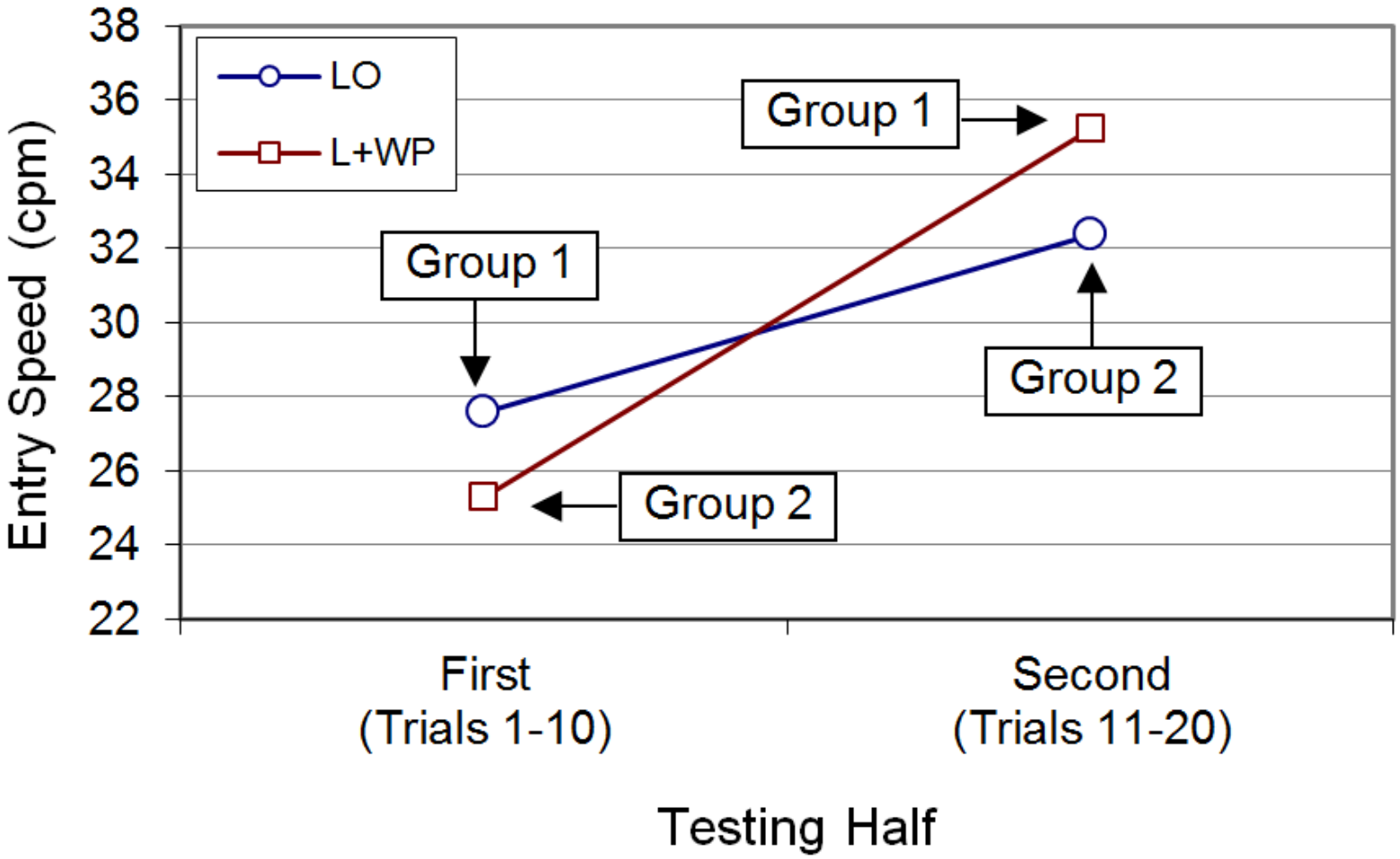
L+WP Keyboard



Testing Half		Group
First (Trials 1-10)	Second (Trials 11-20)	
20.42	27.12	1
22.68	28.39	
23.41	32.50	
25.22	32.12	
26.62	35.94	
28.82	37.66	
30.38	39.07	
31.66	35.64	
32.11	42.76	
34.31	41.06	
19.47	24.97	2
19.42	27.27	
22.05	29.34	
23.03	31.45	
24.82	33.46	
26.53	33.08	
28.59	34.30	
26.78	35.82	
31.09	36.57	
31.07	37.43	

= LO

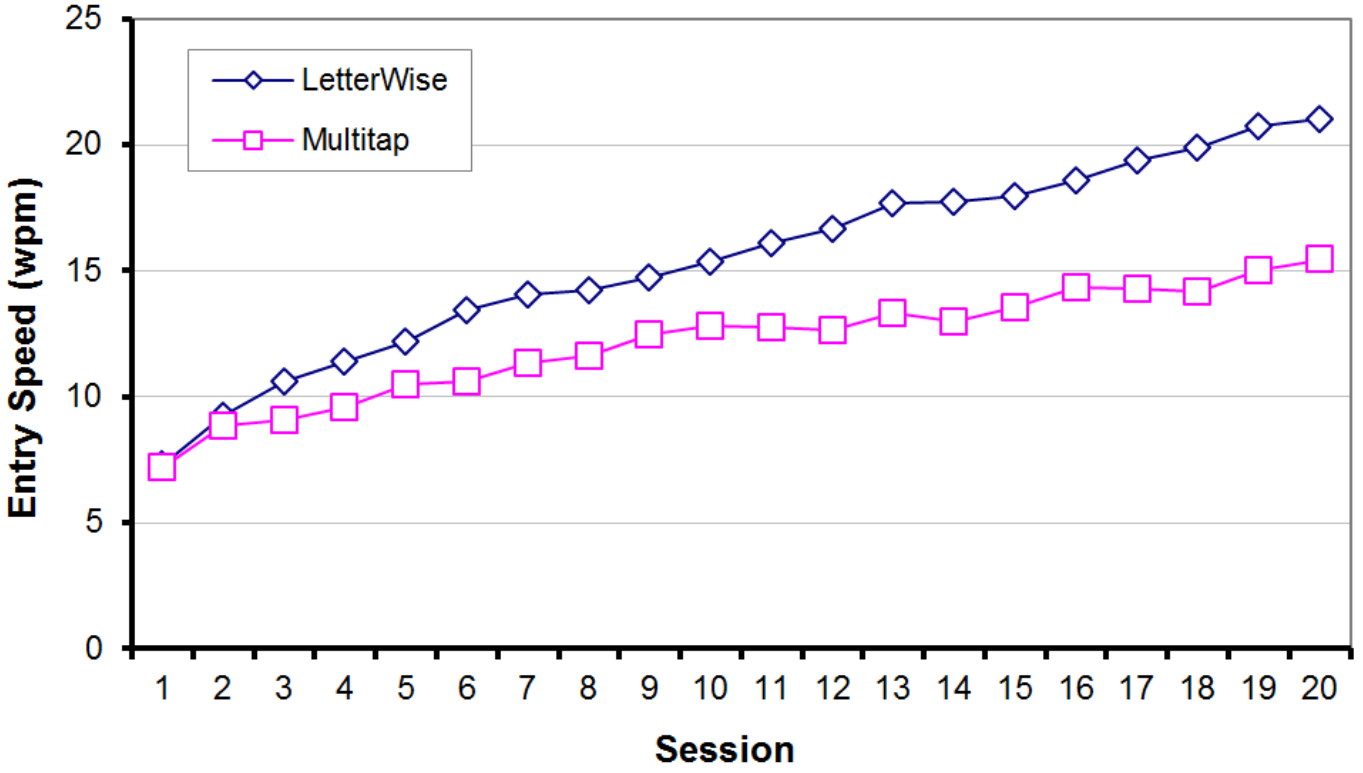

Entry Speed (cpm)
45
40
35
30
25
20
LO
L+WP
Keyboard
First
Second
Testing Half
1
2
Group


38
36
34
32
30
28
26
24
22
Entry Speed (cpm)
LO
L+WP
Group 1
Group 1
Group 2
Group 2
First
(Trials 1-10)
Second
(Trials 11-20)
Testing Half

Longitudinal Studies

- Sometimes instead of "balancing out" learning effects, the research seeks to promote and investigate learning
- If so, a *longitudinal study* is conducted
- "Practice" is the IV
- Participants are practiced over a prolonged period of time
- Practice units: blocks, sessions, hours, days, etc.
- Example on next slide

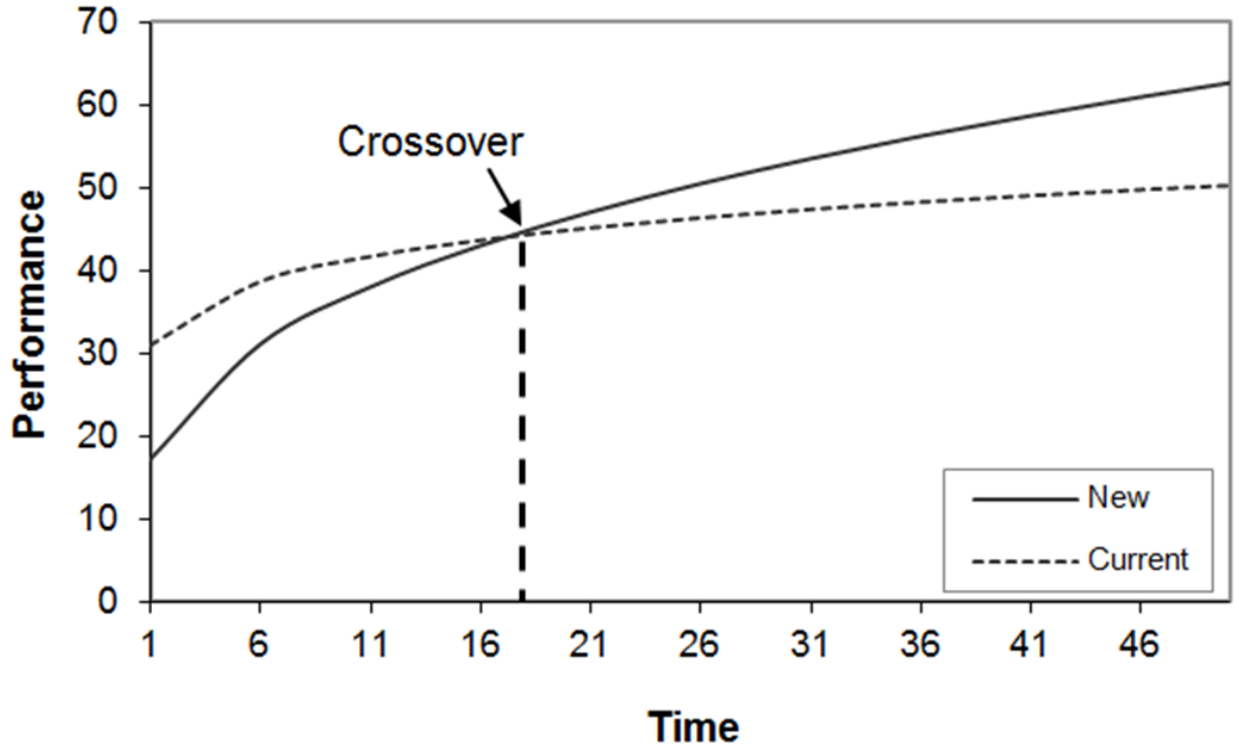
Longitudinal Study – Results[1]



[1] MacKenzie, I. S., Kober, H., Smith, D., Jones, T., & Skepner, E. (2001). LetterWise: Prefix-based disambiguation for mobile text entry. *Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2001*, 111-120, New York: ACM.

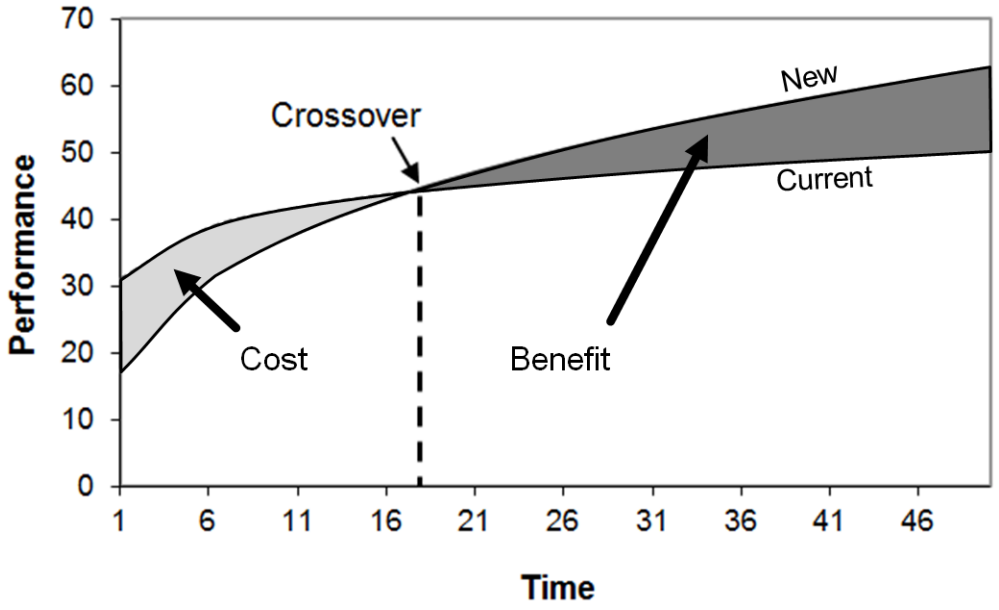
The New vs. The Old

- Sometimes a new technique will initially perform poorly in comparison to an established technique
- A longitudinal study will determine if a crossover point occurs and, if so, after how much practice (see below)



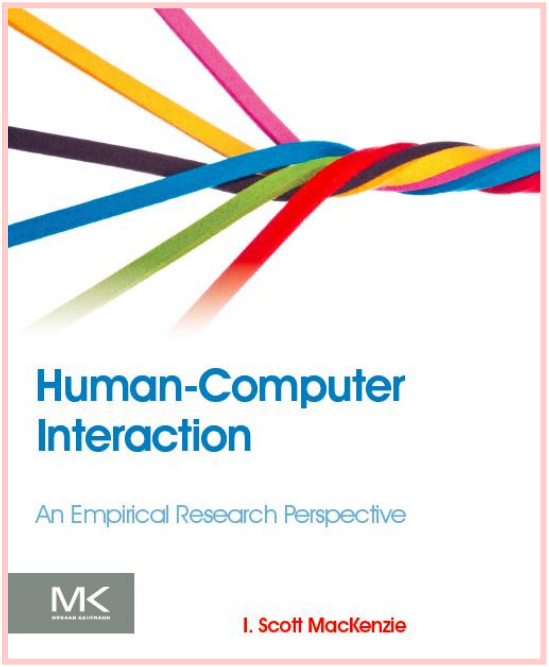
Cost-Benefit Trade-offs

- New, improved techniques sometimes languish
- Evidently, the benefit in the new technique is insufficient to overcome the cost in learning it (see below)



- Of course, there are other issues (see **HCI:ERP** for discussion)

Thank You


Human-Computer Interaction
An Empirical Research Perspective
MK
I. Scott MacKenzie