Empirical studies

design and evaluation

SAN 2018/19

Department of Computer Graphics and Interaction,
Czech Technical University in Prague

DCGI

https://cw.fel.cvut.cz/wiki/courses/b4m36san/start

Introduction

Historical Context

- 1940s first computers
- 1980s SIGCHI formed
- 1940s – 1980s?
  - No users, only engineers, computer scientists
  - Computers were big, expensive, inaccessible

Historical Context

1940

1945 – Vannevar Bush publishes "As We May Think" in *The Atlantic Monthly*

1950

1960

1962 – Ivan Sutherland develops *Sketchpad*

1963 – Douglas Engelbart invents the computer mouse

1970

1980

1981 – Xerox *Star* launched

1982 – ACM SIGCHI formed

1983 – Card, Moran, and Newell publish *The Psychology of Human-Computer Interaction*

1984 – Apple *Macintosh* launched

1990

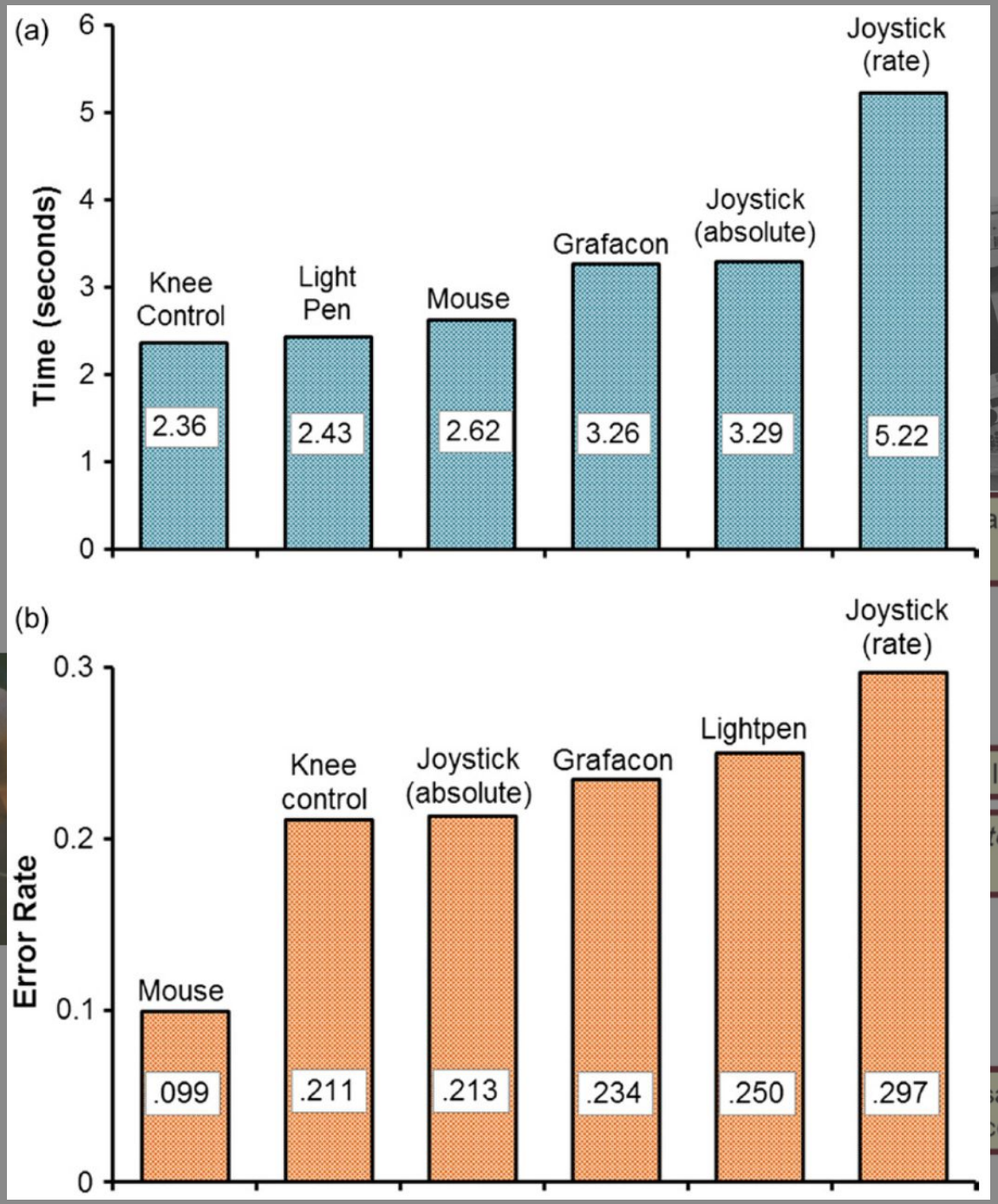
2000

2007 – 25th Anniversary of "CHI", the SIGCHI annual conference

2010

MacKenzie 2013

Historical



and develops

l formed

tosh launched

sary of "CHI", conference

MacKenzie 2013

Historical Context

1945 – Vannevar Bush publishes "As We May Think" in *The Atlantic Monthly*

1940

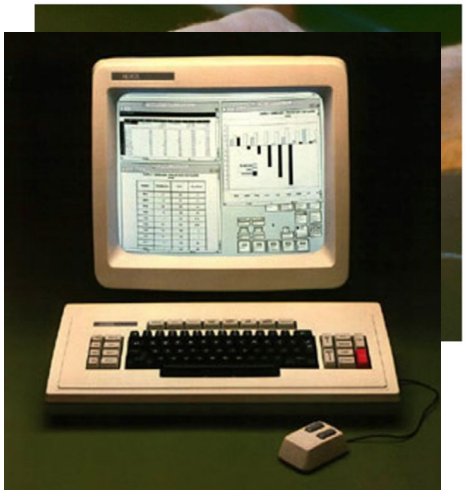
1950

1960

1962 – Ivan Sutherland develops *Sketchpad*

1963 – Douglas Engelbart invents the computer mouse

1970



1981 – Xerox *Star* launched

1980

1982 – ACM SIGCHI formed

1983 – Card, Moran, and Newell publish *The Psychology of Human-Computer Interaction*

1984 – Apple *Macintosh* launched

1990

2000

2007 – 25th Anniversary of "CHI", the SIGCHI annual conference

2010

MacKenzie 2013

Historical Context

1940

1945 – Vannevar Bush publishes "As We May Think" in *The Atlantic Monthly*

1950

1960

1962 – Ivan Sutherland develops *Sketchpad*

1963 – Douglas Engelbart invents the computer mouse

1970

1980

1981 – Xerox *Star* launched

1982 – ACM SIGCHI formed

1983 – Card, Moran, and Newell publish *The Psychology of Human-Computer Interaction*

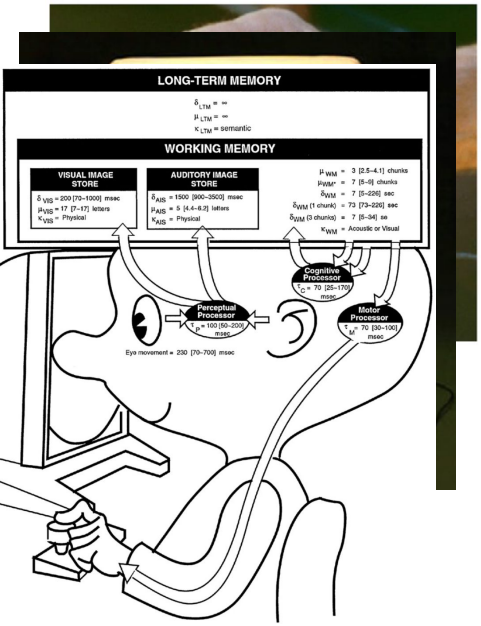
1984 – Apple *Macintosh* launched

1990

2000

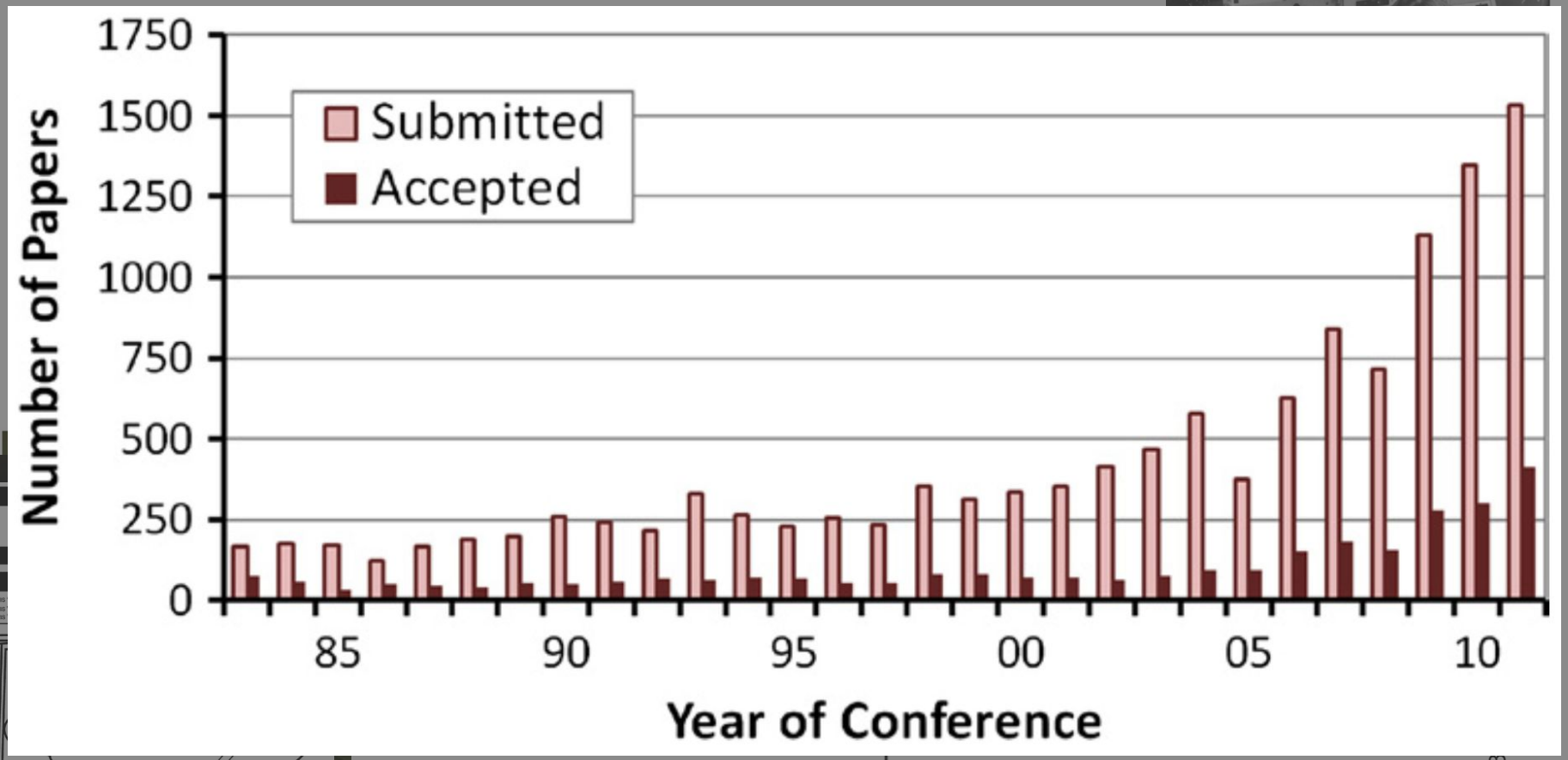
2007 – 25th Anniversary of "CHI", the SIGCHI annual conference

2010



MacKenzie 2013

Historical Context



2000

2007 – 25th Anniversary of "CHI", the SIGCHI annual conference

2010

MacKenzie 2013

Human Factors

Human Factors

Humans are complicated – Computers are simple

- Old, female, male, experts, novices, left- handed, right-handed, English-speaking, Chinese-speaking, from the north, from the south, tall, short, strong, weak, fast, slow, able-bodied, disabled, sighted, blind, motivated, lazy, creative, bland, tired, alert, …
- Humans are never precise

Time scale of human actions

- workplace habits, g
  usage patterns, so
  networking, online
  privacy, media spa
  theory, …
- web navigation, us
  strategies, user-ce
  collaborative comp
  ubiquitous comput
  navigation, …
- selection technique
  auditory feedback,
  gestural input, …

Scale (sec)	Time Units	System	World (theory)
10^7	Months		**SOCIAL BAND**
10^6	Weeks		
10^5	Days		
10^4	Hours	Task	**RATIONAL BAND**
10^3	10 min	Task	
10^2	Minutes	Task	
10^1	10 sec	Unit task	**COGNITIVE BAND**
10^0	1 sec	Operations	
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	**BIOLOGICAL BAND**
10^{-3}	1 ms	Neuron	
10^{-4}	100 μs	Organelle	

Newell 1990

Time scale of human actions

- workplace habits, groupware usage patterns, social networking, online dating, privacy, media spaces, design theory, …
- web navigation, user search strategies, user-centered design, collaborative computing, ubiquitous computing, social navigation, …
- selection techniques, force or auditory feedback, text entry, gestural input, …

Scale (sec)	Time Units	System	World (theory)
10^{7}	Months		**SOCIAL BAND**
10^{6}	Weeks		
10^{5}	Days		
10^{4}	Hours	Task	**RATIONAL BAND**
10^{3}	10 min	Task	
10^{2}	Minutes	Task	
10^{1}	10 sec	Unit task	**COGNITIVE BAND**
10^{0}	1 sec	Operations	
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	**BIOLOGICAL BAND**
10^{-3}	1 ms	Neuron	
10^{-4}	100 µs	Organelle	

Newell 1990

Time scale of human actions

Qualitative

Quantitative

- workplace habits, groupware usage patterns, social networking, online dating, privacy, media spaces, design theory, …
- web navigation, user search strategies, user-centered design, collaborative computing, ubiquitous computing, social navigation, …
- selection techniques, force or auditory feedback, text entry, gestural input, …

Scale (sec)	Time Units	System	World (theory)
10^7	Months		**SOCIAL BAND**
10^6	Weeks		
10^5	Days		
10^4	Hours	Task	**RATIONAL BAND**
10^3	10 min	Task	
10^2	Minutes	Task	
10^1	10 sec	Unit task	**COGNITIVE BAND**
10^0	1 sec	Operations	
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	**BIOLOGICAL BAND**
10^{-3}	1 ms	Neuron	
10^{-4}	100 μs	Organelle	

Newell 1990

Sensors

Vision

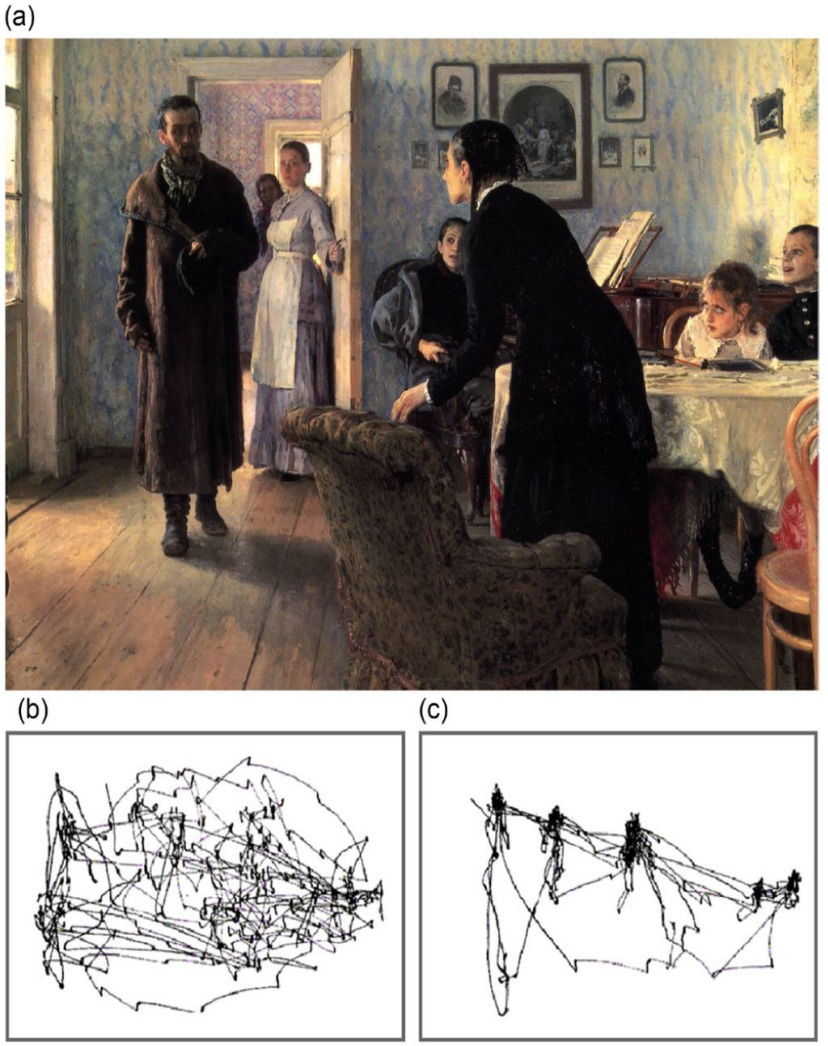
- Intensity, Fixations, Saccades

Hearing

- Loudness, Pitch, Timbre

Touch

- Position, Texture, Temperature, Movement, Resistance



Tatler 2010

(a) Scene. (b) Task: Remember the position of the people and objects in the room. (c) Task: Estimate the ages of the people

Sensors

Vision

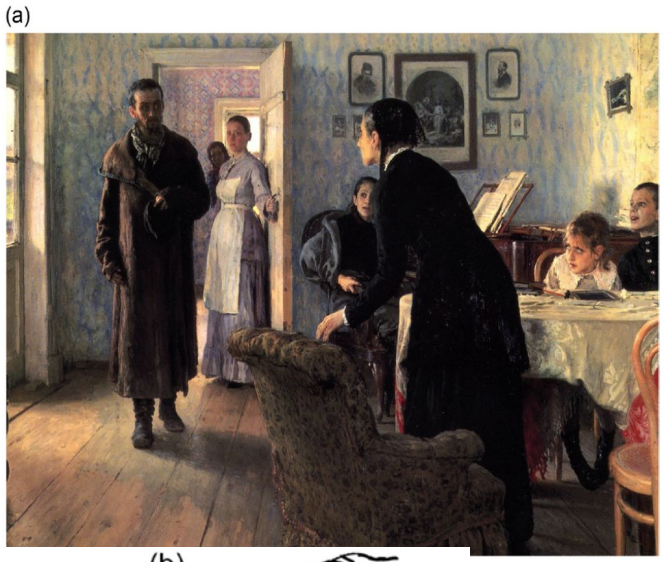
- Intensity, Fixations, Saccades

Hearing

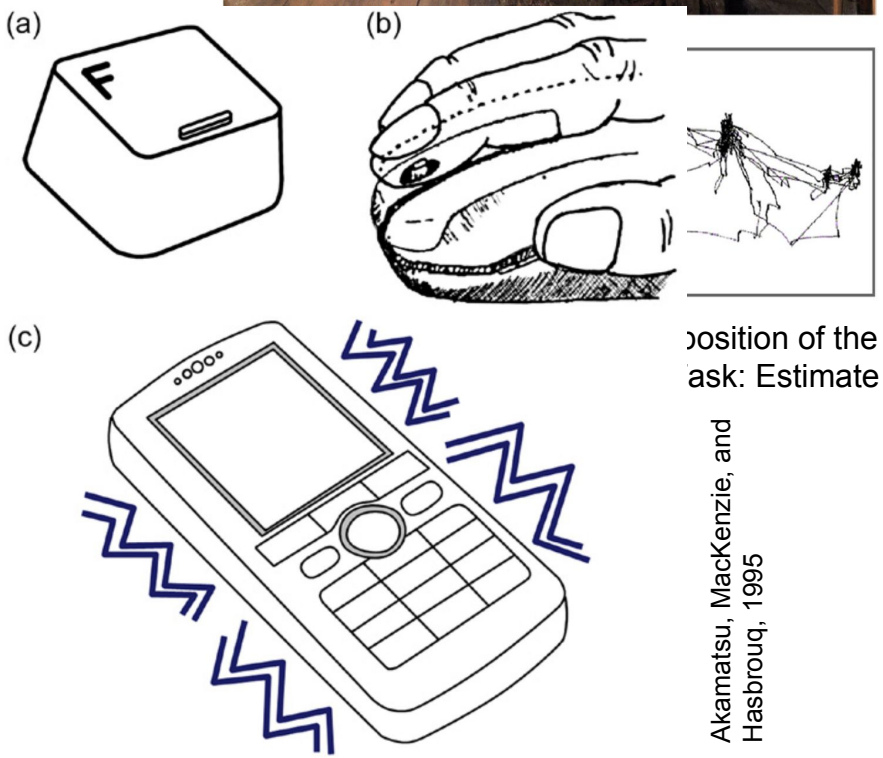
- Loudness, Pitch, Timbre

Touch

- Position, Texture, Temperature, Movement, Resistance



Tatler 2010



osition of the
ask: Estimate

Akamatsu, MacKenzie, and Hasbrouq, 1995

(a) Identifier on key top. (b) Solenoid-driven pin under the index finger. (c) Vibration signals an in-coming call

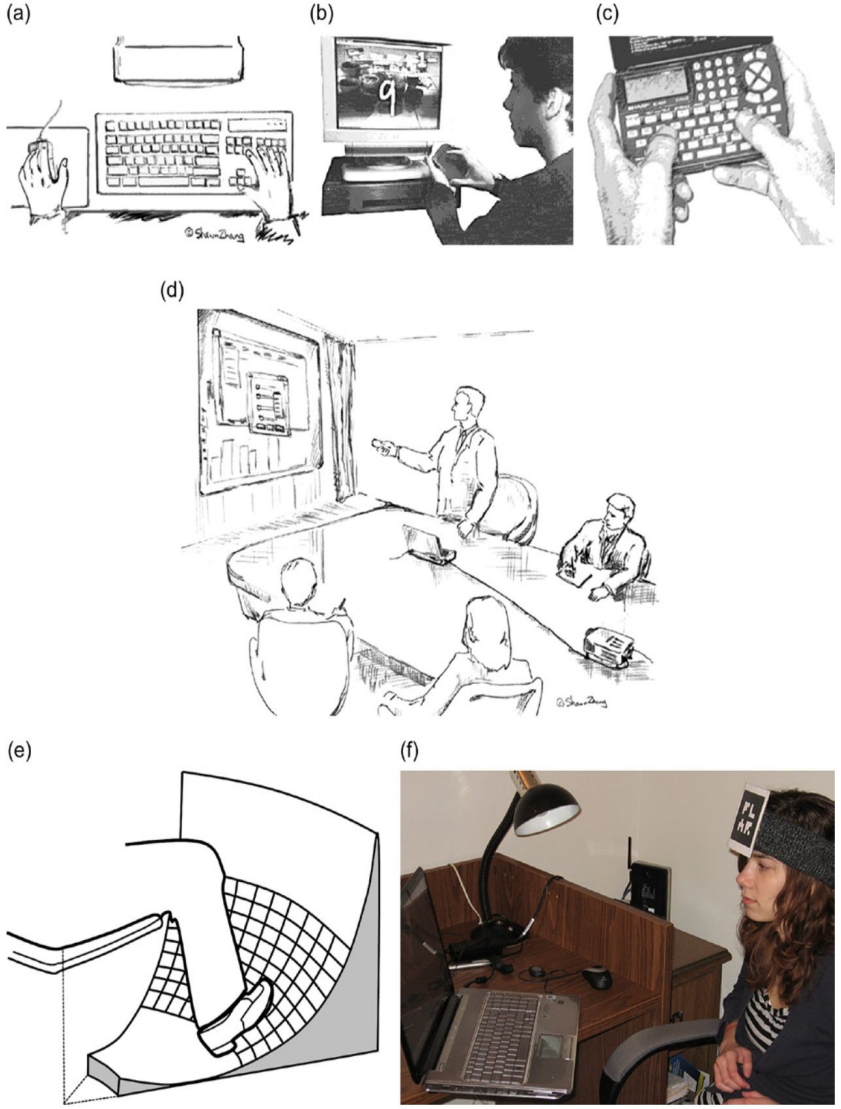
Responders

Limbs

Voice

Eyes

Taste and smell



use of the limbs in HCI: (a) Hands. (b) Fingers. (c) Thumbs. (d) Arms. (e) Feet. (f) Head.

a and d courtesy of Shawn Zhang; e, adapted from Pearson and Weiser, 1986, MacKenzie 2013

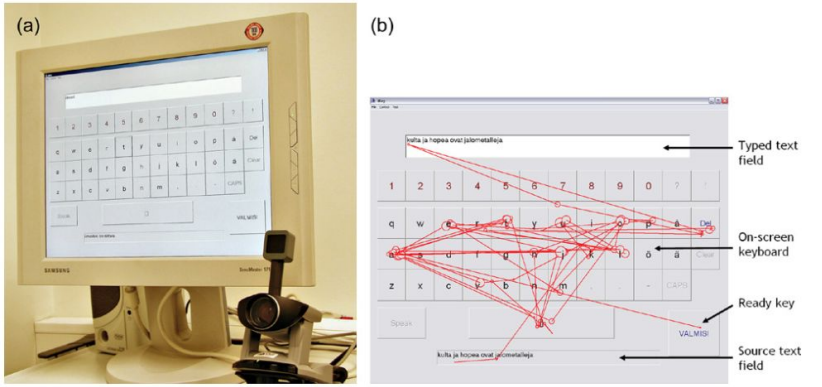
Responders

Limbs

Voice

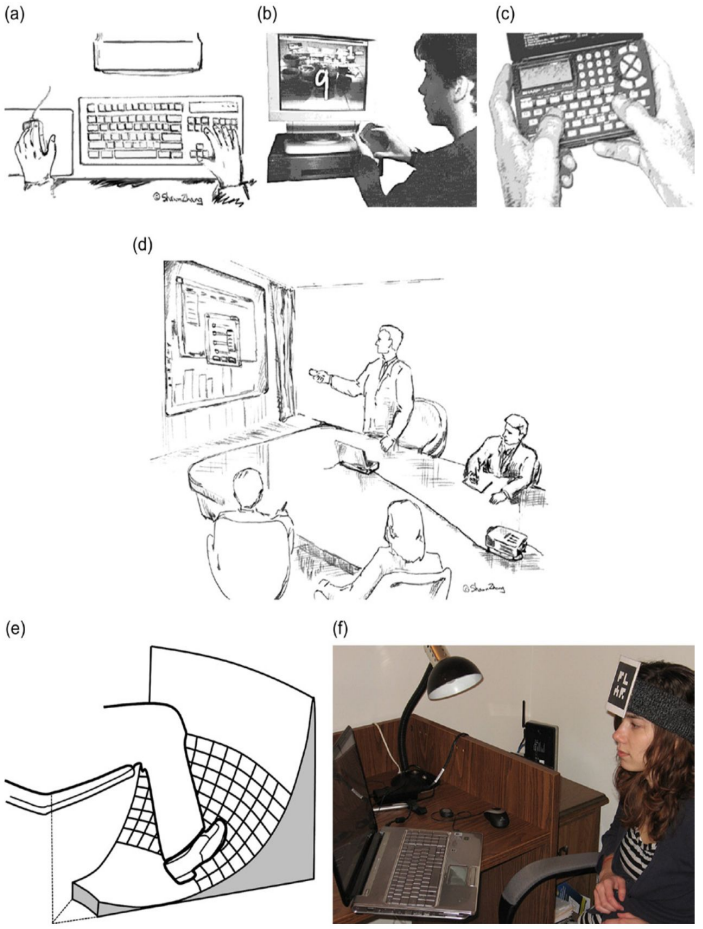
Eyes

Taste and smell



use of the limbs in HCI: (a) Hands. (b) Fingers. (c) Thumbs. (d) Arms. (e) Feet. (f) Head.

Majaranta et al., 2006



use of the limbs in HCI: (a) Hands. (b) Fingers. (c) Thumbs. (d) Arms. (e) Feet. (f) Head.

a and d courtesy of Shawn Zhang; e, adapted from Pearson and Weiser, 1986, MacKenzie 2013

Brain

Cognition

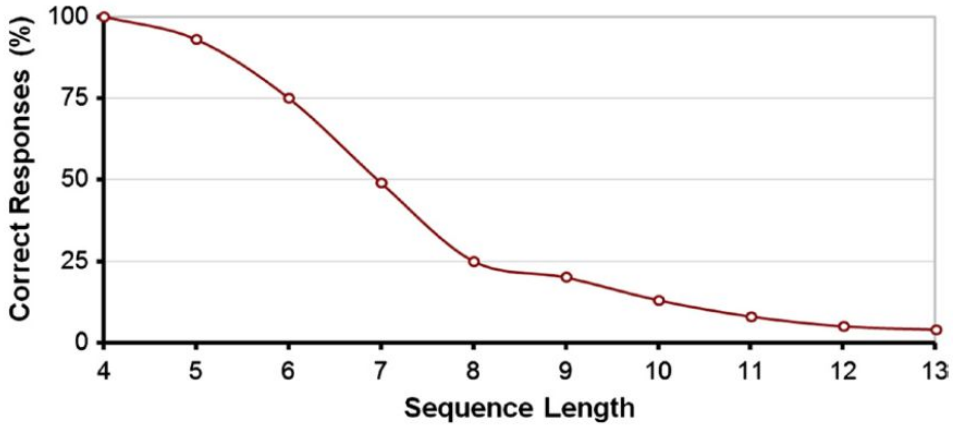
- Thinking, reasoning, and deciding

Memory

- Long-term vs short-term (working)

Language

- Corpus, redundancy, entropy



Short-term Memory

Brain

Cognition

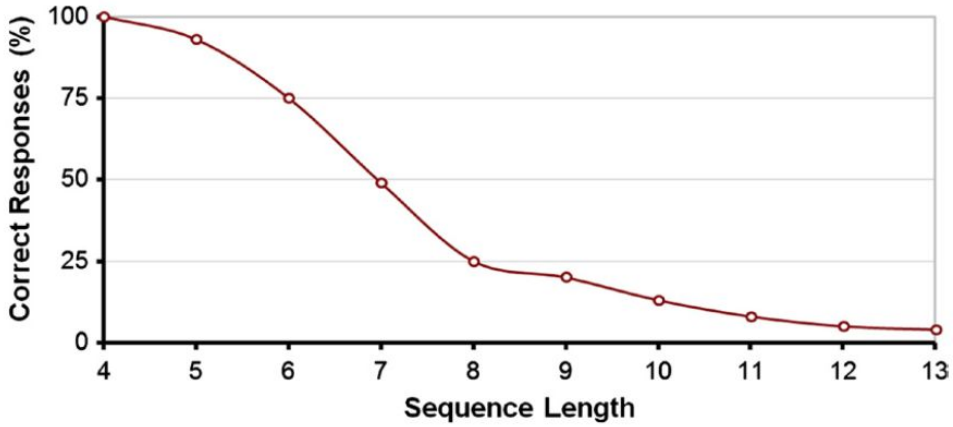
- Thinking, reasoning, and deciding

Memory

- Long-term vs short-term (working)

Language

- Corpus, redundancy, entropy



Short-term Memory

```
THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
----ROO------NOT-V-----I------SM----OB----

READING LAMP ON THE DESK SHED GLOW ON
REA----------O------D----SHED-GLO--0-

POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
P-L-S-----O---BU--L-S--O-------SH----RE--C-----
```

Shannon's letter-guessing experiment.

Human Performance

Reaction time

- Delay between the occurrence of a stimulus and the initiation of a response

Visual search

- Linear relation to number of items

Skilled behaviour

- Performance improves through training

Attention

- Concentrating on a discrete aspect of information, while ignoring other perceivable information

Error

- Error is a discrete event in a task, or trial, where the outcome is incorrect

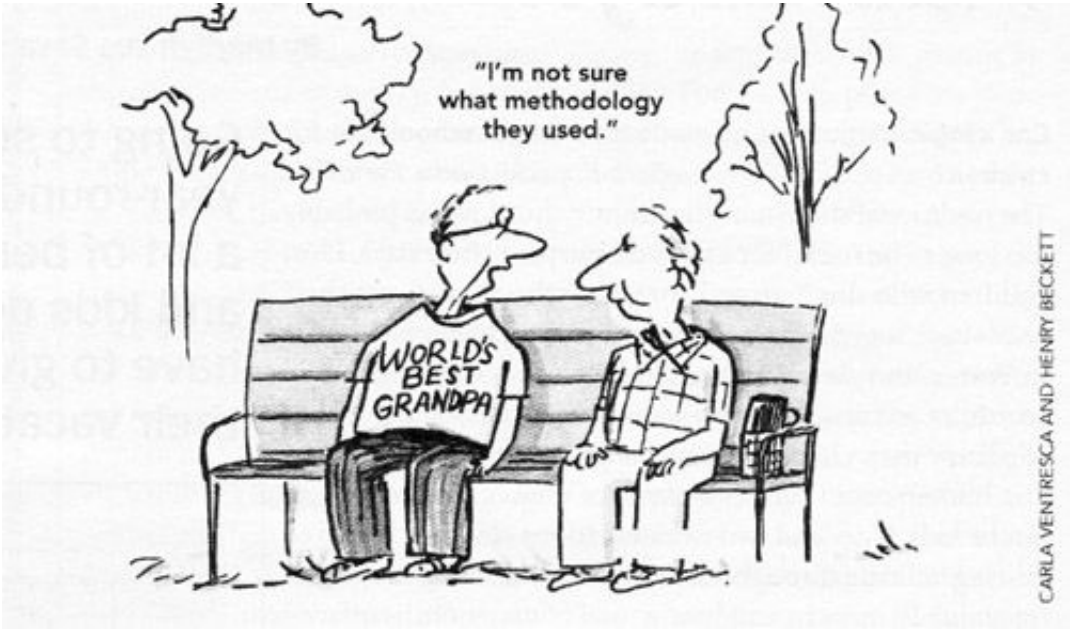
Research Methods

Research methods

Observational method

Experimental method

Correlational method



Observational method

- Interviews, field investigations, contextual inquiries, case studies, field studies, focus groups, think aloud protocols, storytelling, walkthroughs, cultural probes, …
- Focus on human thought, feeling, attitude, emotion, passion, sensation, re ection, expression, sentiment, opinion, mood, outlook, manner, style, approach, strategy, …
- Qualitative rather than quantitative
- Achieve relevance while sacrificing precision

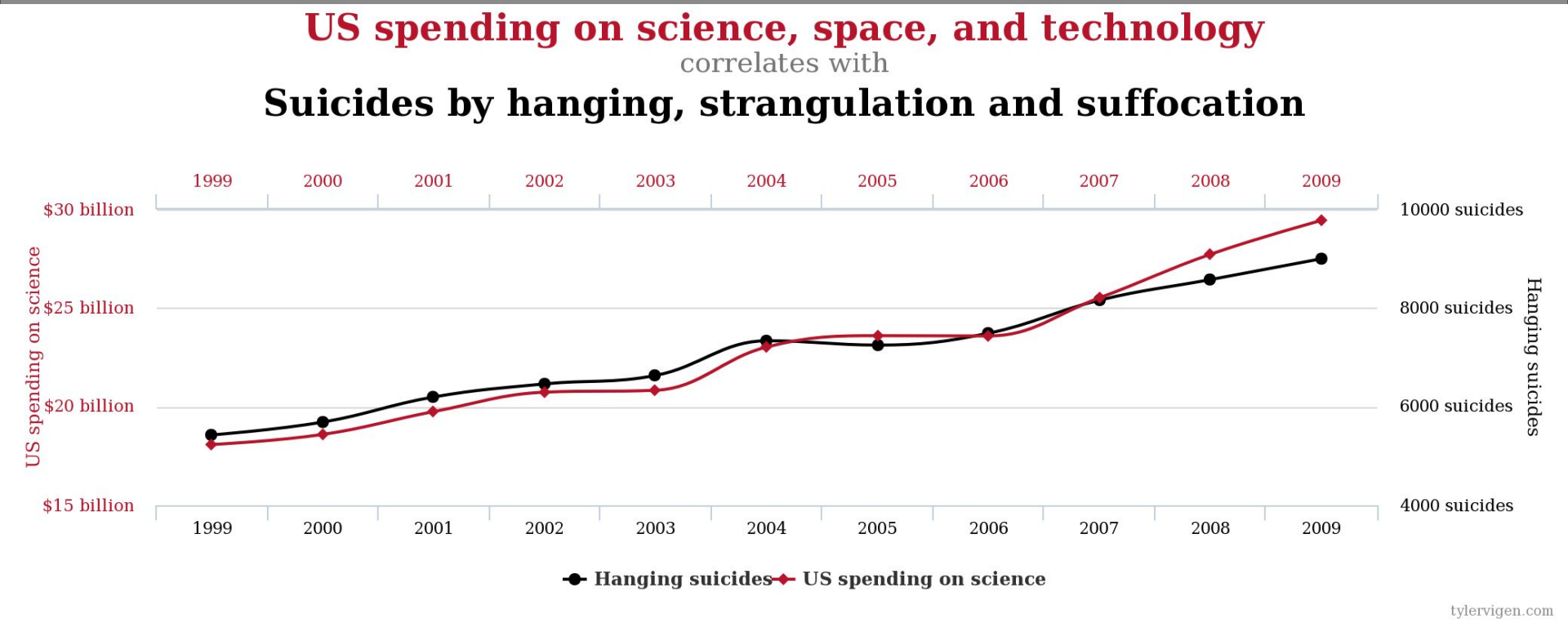
Experimental method

- Knowledge is acquired through controlled experiments conducted in laboratory settings
- At least **two variables** – a manipulated (independent) variable and a response (dependent) variable
    - systematically exposing participants to different configurations of the interface or interaction technique
- Task completion time, …
- Control inherent in the methodology brings precision
- Allows conclusion to be drawn from the data and analyses
    - Unlike from the other two methods
    - We change manipulated variable and observer change in response variable

Correlational method

- Looking for relations between variables
- Characterized by quantification since the magnitude of variables must be ascertained (e.g., age, income, number of privacy settings)
- For nominal-scale variables, categories are established (e.g., personality type, gender)
- The data are collected through a variety of methods, such as observation, interviews, on-line surveys, questionnaires, or measurement
- Correlational methods often accompany experimental methods, if questionnaires are included in the experimental procedure
- Balance between relevance and precision

Correlational method

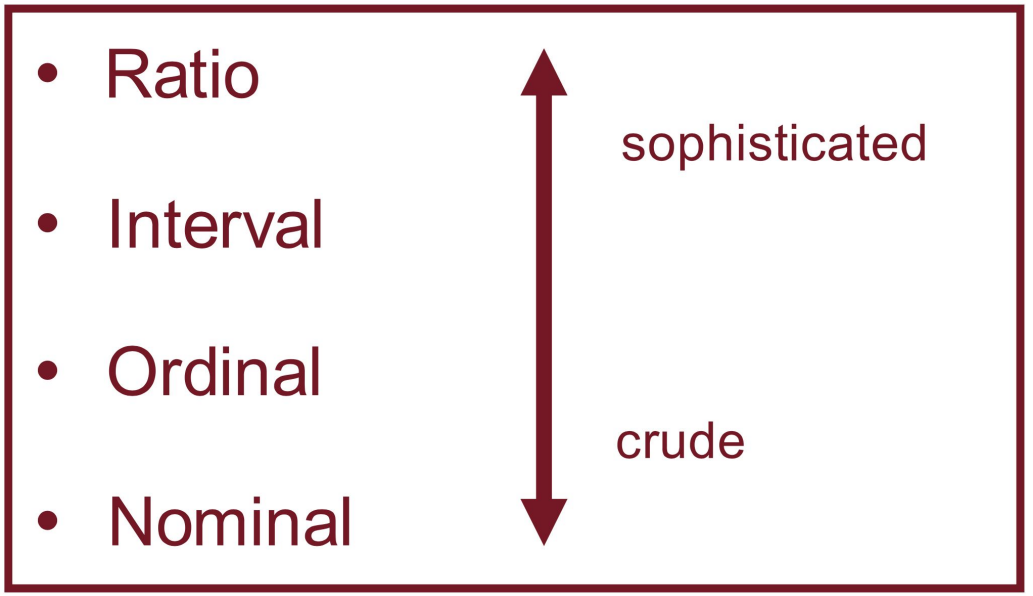


- Balance between relevance and precision

Measurement

Measurement scales

- Nominal, ordinal, interval, ratio
- Nature, limitations, and abilities of each scale determine the sort of information and analyses possible



MacKenzie 2013

Nominal

- Assigning a code to an attribute or a category (it does not need to be a number)
- Often used with frequencies or counts

P02	F	BHAL	L	4
P06	F	AHBL	C	4
P07	F	ALBH	C	4
P08	F	BHAL	C	5
P09	F	BLAH	C	5
P10	F	AHBL	C	5
P11	M			
P13	M			
P14	M			
P15	F			
P16	F			
P18	M			
P19	F			
P20	M			

Gender	Mobile Phone Usage		Total	%
	Not Using	Using		
Male	683	98	781	51.1%
Female	644	102	746	48.9%
Total	1327	200	1527	
%	86.9%	13.1%		

Ordinal

- Ordinal scale measurements provide an order or ranking to an attribute
- Interval is not intrinsically equal between successive points on the scale
- Comparisons of greater than or less than are possible
- It is not valid to compute the mean of ordinal data

How many email messages do you receive each day?

1. None (I don't use email)
2. 1-5 per day
3. 6-25 per day
4. 26-100 per day
5. More than 100 per day

MacKenzie 2013

Interval

- Interval data have equal distances between adjacent values
- There is no absolute zero
- E.g. temperature measured on a scale
  - It is meaningful to compute the mean of interval data
  - Ratios of interval data are not meaningful – one cannot say that 20°C is twice as warm as 10° C

Please indicate your level of agreement with the following statements.

	Strongly disagree	Mildly disagree	Neutral	Mildly agree	Strongly agree
It is safe to talk on a mobile phone while driving.	1	2	3	4	5
It is safe to read a text message on a mobile phone while driving.	1	2	3	4	5
It is safe to compose a text message on a mobile phone while driving.	1	2	3	4	5

Ratio

- Ratio data have an absolute zero and support a many of calculations to summarize, compare, and test the data
- In HCI, the most common ratio-scale measurement is time
  - Generally, all physical measurements
- Another common ratio-scale measurement is count
  - Count is improved through **normalization**; that is, expressing the value as a count **per something**
- Errors normalized as “error rates (%)”
  - E.g. number of errors/number of trials*100 – number of incorrectly entered characters/total number of characters times 100

Research question?

Research Question

Research is conducted to answer (**and raise**) questions about new or existing user interfaces or interaction techniques

Often the questions contains the relationship between two variables:

- One variable is a circumstance or condition that is manipulated – interface property
- The other is an observed and measured behavioral response – task performance

Research Question

Questions about the UI or interaction techniques:

- Is it viable?
- Is it as good as or better than current practice?
- What are its strengths and weaknesses?
- Which of several alternatives is best?

Research Question

Questions about the UI or interaction techniques:

- Is it viable?
- Is it as good as or better than current practice?
- What are its strengths and weaknesses?
- Which of several alternatives is best?

Relevant, but not testable!

Research Question

Example, questions about new technique comparing to qwerty software keyboard (QSK).

- Is the new technique any good?
- Is the new technique better than QSK?
- Is the new technique faster than QSK?
- Is the new technique faster than QSK after a bit of practice?
- Is the measured entry speed (in words per minute) higher for the new technique than for a QSK after one hour of use?

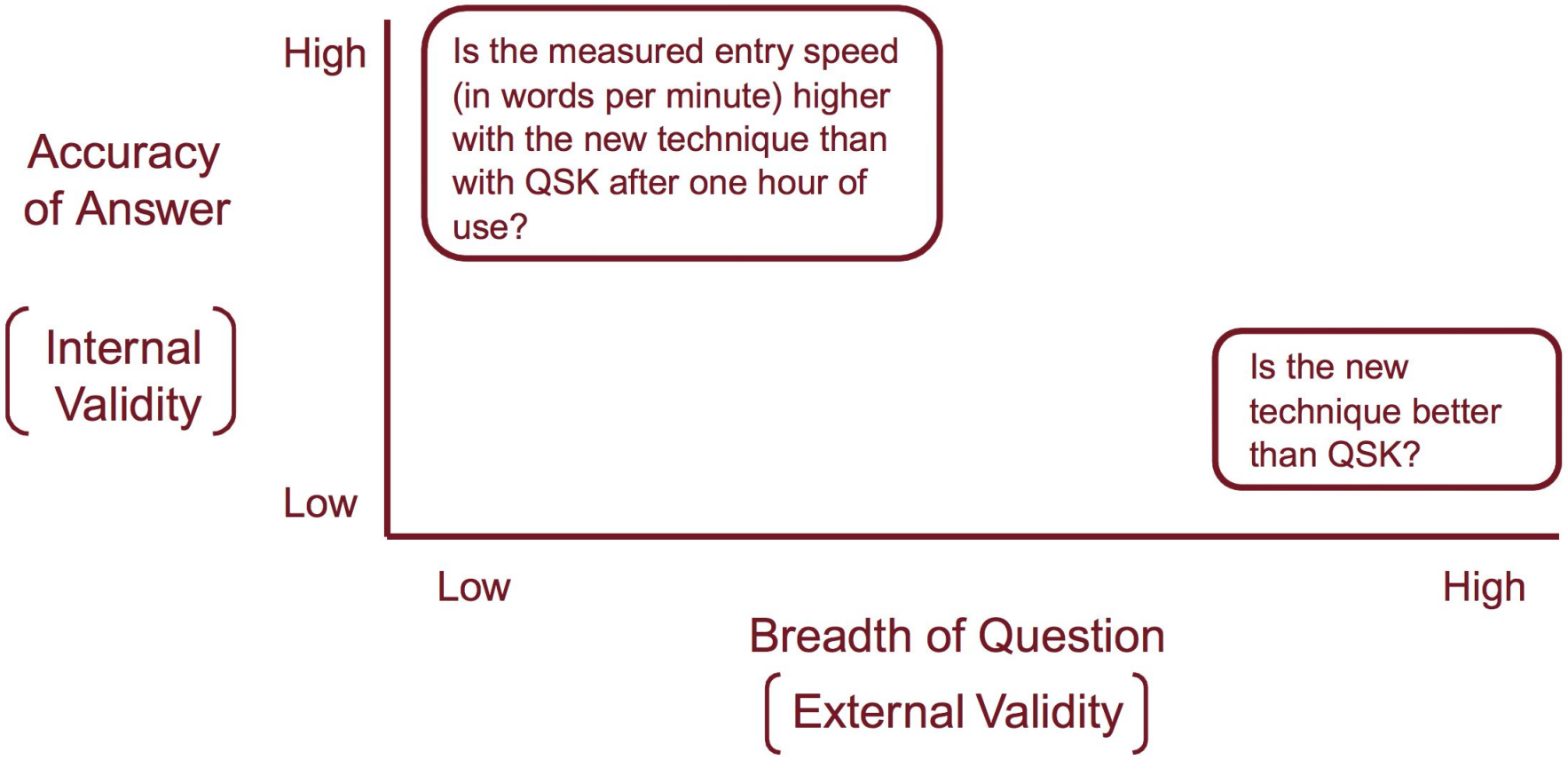
Research Question

Example, questions about new technique comparing to qwerty software keyboard (QSK).

- Is the new technique any good?
- Is the new technique better than QSK?
- Is the new technique faster than QSK?
- Is the new technique faster than QSK after a bit of practice?
- **Is the measured entry speed (in words per minute) higher for the new technique than for a QSK after one hour of use?**

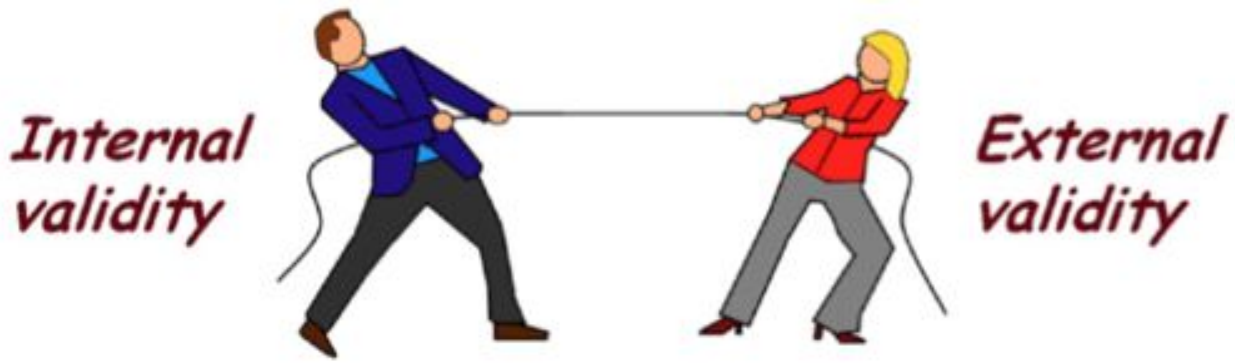
More focused

Internal vs. External Validity



MacKenzie 2013

Internal vs. External Validity



Internal Validity

low in breadth (that's bad!) yet answerable with high accuracy (that's good!)> we can craft a methodology to answer it through observation and measurement

External Validity

high in breadth (that's good!) yet answerable with low accuracy (that's bad)> we lack a methodology to observe and measure "better than"

Variability

- People exhibit variability in their actions
- Variability person per person, but also person per task

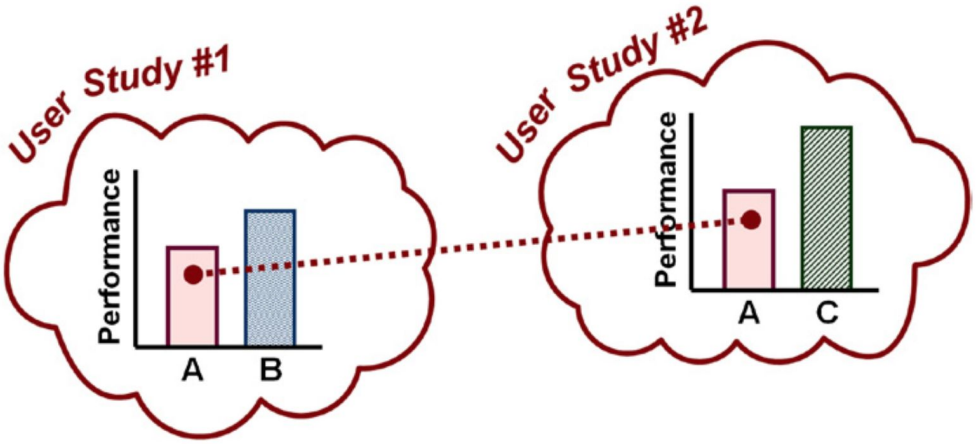
The result is always different!

This variability affects the confidence with which we can answer research questions.

Designing HCI Experiments

Comparative evaluations

- A new UI or interaction technique evaluated on its own is questionable
- It need to be compared to other design (baseline condition)
  - if the new one is faster, more accurate, less confusing, more preferred by users, … than the baseline condition
- The testable research questions are crafted as comparisons



MacKenzie 2013

Including a baseline condition serves as a check on the methodology and facilitates the comparison of results between user studies.

Experiment design

Process of bringing together all the pieces necessary to test hypotheses on a user interface or interaction technique

- Variables
- Tasks and procedure
- Participants

Independent variables

An independent variable (factor) is a circumstance or characteristic that is manipulated or systematically controlled to a change in a human response while the user is interacting with a computer.

- Manipulated across multiple levels (at least 2) or test conditions
- It is “independent” because it is independent of participant behavior
- Typically a nominal-scale attribute, often related to a property of an interface
  - Such as device, entry method, feedback modality, selection technique, menu depth, button layout
  - It can bye also characteristic of a human (age, handedness, gender, expertise, …) – naturally occurring attributes, they cannot be manipulated
  - Environment characteristics (room lightning, background noise, …)

Dependent variables

A dependent variable is a measured human behavior

- Typically a ratio-scale human behavior
  - task completion time, error rate, accuracy, number of button clicks, scrolling events, gaze shifts, …
- The “dependent” in dependent variable refers to the variable being dependent on the human
  - measurements depend on what the participant does
- Any observable, measurable aspect of human behavior is a potential dependent variable
  - It is essential to clearly de ne all dependent variables to ensure the research can be replicated

Effects

Main effect vs. interaction effects on dependent variables

- Interaction effects that are three-way or higher are extremely difficult to interpret
- Optimal number of independent variables: **one or two**, three at most

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

MacKenzie 2013

Data collection

It is important to think about hout the measurements will be gathered, stored, organized

- Create experimental software to capture timestamps, key presses, etc.
- When organizing data, think about future analysis
- Pilot testing is crucial
  - To ensure everything work including data collection
  - Perform preliminary analysis

P01	2	1	1	1	1	1	1
P08	1	1	4	1	1	1	1
P13	3	1	2	1	2	1	1
P14	1	1	1	2	1	2	1
P06	2	1	1	1	5	3	1
P02	1	1	2	1	2	1	1
P11	2	1	1	2	1	2	1
P18	1	1	1	1	1	3	1
P10	2	1	4	1	1	2	1

Other variables

Control variables

- Might influence a dependent variable but are not under investigation
  - Room lighting, room temperature, background noise, display size, mouse shape, mouse cursor speed, keyboard angle, chair height, or participant characteristic as vision, handedness

Random variables

- Reduce the variability in the measured behaviors – results that are less generalizable
  - Typically characteristics of the participants, including biometrics (height, weight, hand size), social disposition (nervousness), genetics (gender, IQ)

Variable	Advantage	Disadvantage
Random	Improves external validity by using a variety of situations and people.	Compromises internal validity by introducing additional variability in the measured behaviours.
Control	Improves internal validity since variability due to a controlled circumstance is eliminated	Compromises external validity by limiting responses to specific situations and people.

Other variables

Confounding variables

- Any circumstance or condition that changes systematically with an independent variable is a confounding variable
- Very problematic in research – is the effect due to independent variable or confounding?
- E.g. prior experience, experiment setup (different in conditions), ...

Task and procedure

test conditions

- Independent variable x values
  - Pointing device (mouse, joystick, …)
  - Expressivity of interface (high, low,...)
- Good procedure should contain all combinations of independent variable and their values

Task and procedure

represent and discriminate

- Good task is *representative* of an activity people do with the interface
  - Improves external validity
  - More representative the task, it's likely to include behaviors unrelated to UI or technique tested
- Good task can *discriminate* the test conditions
  - Attune to the points of differentiation between test conditions

The experimental procedure includes the task but also the instructions, demonstration, or practice given to the participants

Participants

Select participants from the same population to whom to results apply

Use *sufficient* number of participants

- Increasing the number of participants increases the likelihood of achieving statistically significant results
- If not enough participants are used, statistical significance may fail to appear
- Large number of participants: statistically significant results for a difference of no practical significance
- What to do? Search similar research and use similar number of participants
    - Or you can use a priori power analysis to calculate number of participants needed (next lecture)

Participants

Recruiting

- Solicited personally, by email, snowball method, notice on a wall, ...
- Typically from a pool of individuals (members of workplace, students, organizations)
    - Ideally randomly from population
- Screener used to identify population
    - Short questionnaire about demographic data, experience with computer, anything relevant

Participants are required to sign a consent form prior to testing

Within-subjects and between-subjects

Within-subjects is also called repeated measures, because the measurements on **each test condition** are repeated **for each participant.**

- Less participants, more tests per participants
- Variance due to participants' predispositions approximately the same across test conditions
- No need for balancing the groups of participants

For **between-subjects** design, a separate group of participants is used for each test condition.

- More participants, 1 test per participant
- No interference between test conditions (participants cannot "unlearn" one condition before testing on another condition in within-subject)

In **mixed design** some factors are within-subject (blocks) some between-subject (handedness)

Order effects

In within-subject designs participants are tested with one condition, then another, another, …

- This can result in interference between test conditions
  - Learning (practice) effect, fatigue effect – order (sequence) effect in general
- Confounding influence of practice seriously compromises the comparison

Counterbalancing, and latin squares

Counterbalancing

Simplest case 1 factor, 2 levels (A, B), within-subject experiment participants are divided into two groups, 12 participants:

- 6 in one group order A, B
- 6 in the other group order of conditions B, A

This is the simplest case of *Latin square*

- *n* × *n* table filled with *n* different symbols positioned such that each symbol occurs exactly once in each row and each column

(a)

A	B
B	A

(b)

A	B	C
B	C	A
C	A	B

(c)

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

(d)

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

Counterbalancing, and latin squares

Balanced Latin squares

In **balanced Latin squares** where each condition precedes and follows other conditions an **equal number of times**

- Number of levels of the factor must divide equally into the number of participants. If a factor has three levels, then the experiment requires multiple-of-3 participants

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

4x4 unbalanced Latin square

(a)

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

(b)

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

Balanced Latin squares (a) 4 × 4. (b) 6 × 6.

Group effects and asymmetric skill transfer

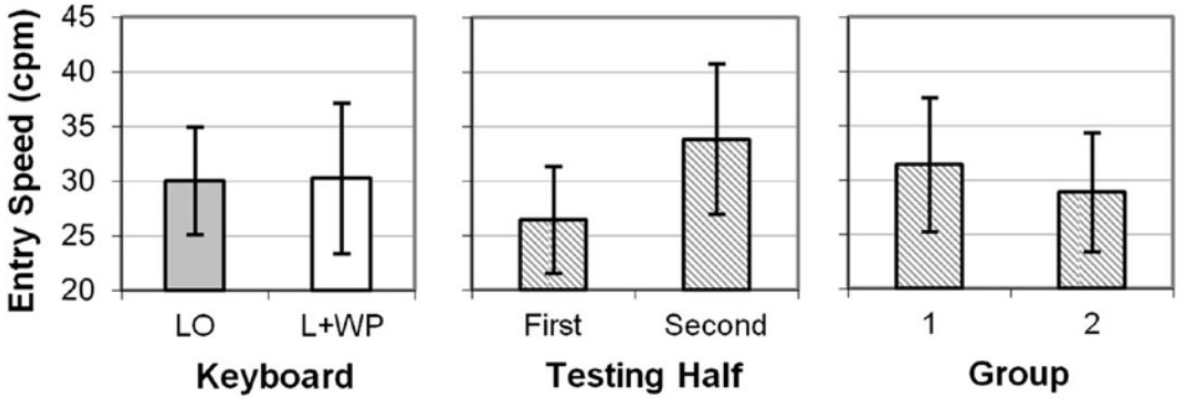
The advantage due to practice for a condition tested later in the experiment is offset equally by the disadvantage when the same condition is tested earlier in the experiment.

- There are occasions where different effects appear for one order (e.g., A→B) compared to another (e.g., B→A)
- Group effect is typically caused by asymmetrical skill transfer
  - Different amount of improvement depending on the order of testing

Asymmetric skill transfer

Skills from first condition transfers to next condition e.g. unskilled/untrained participants

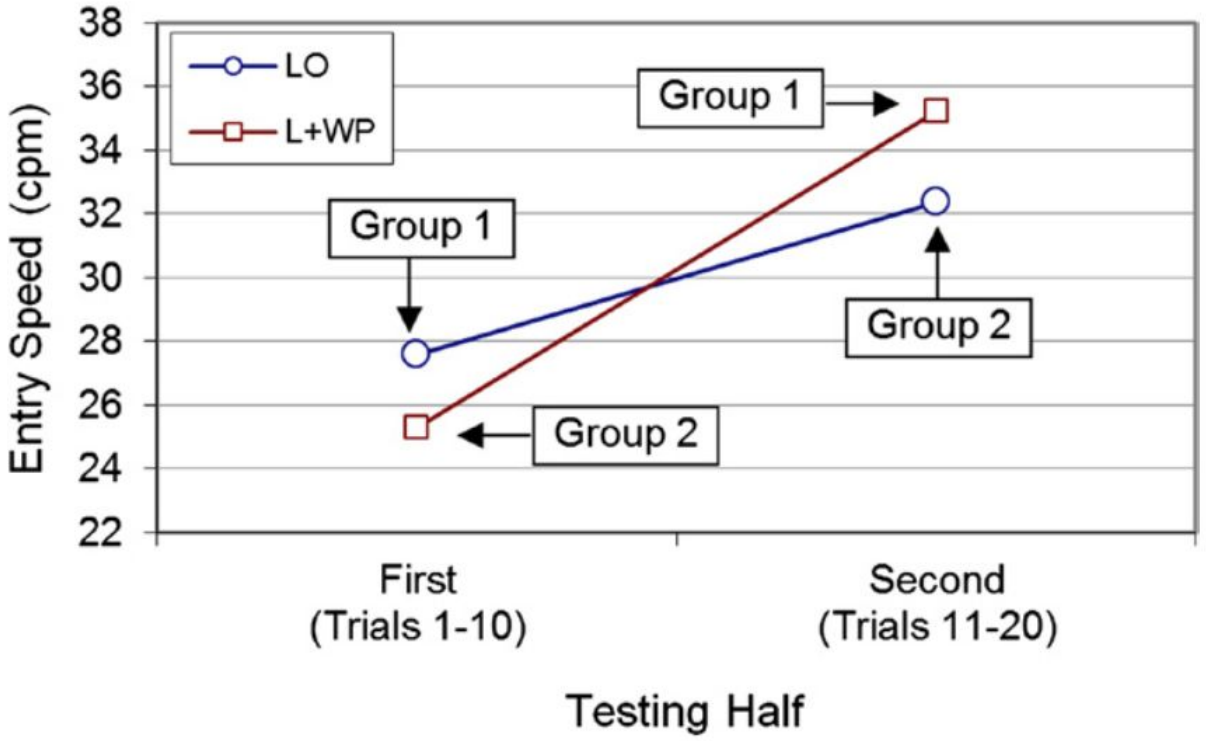
- This can be prevented either by between-subject design, or long enough training in within-subject design



Asymmetric skill transfer

Skills from first condition transfers to next condition e.g. unskilled/untrained participants

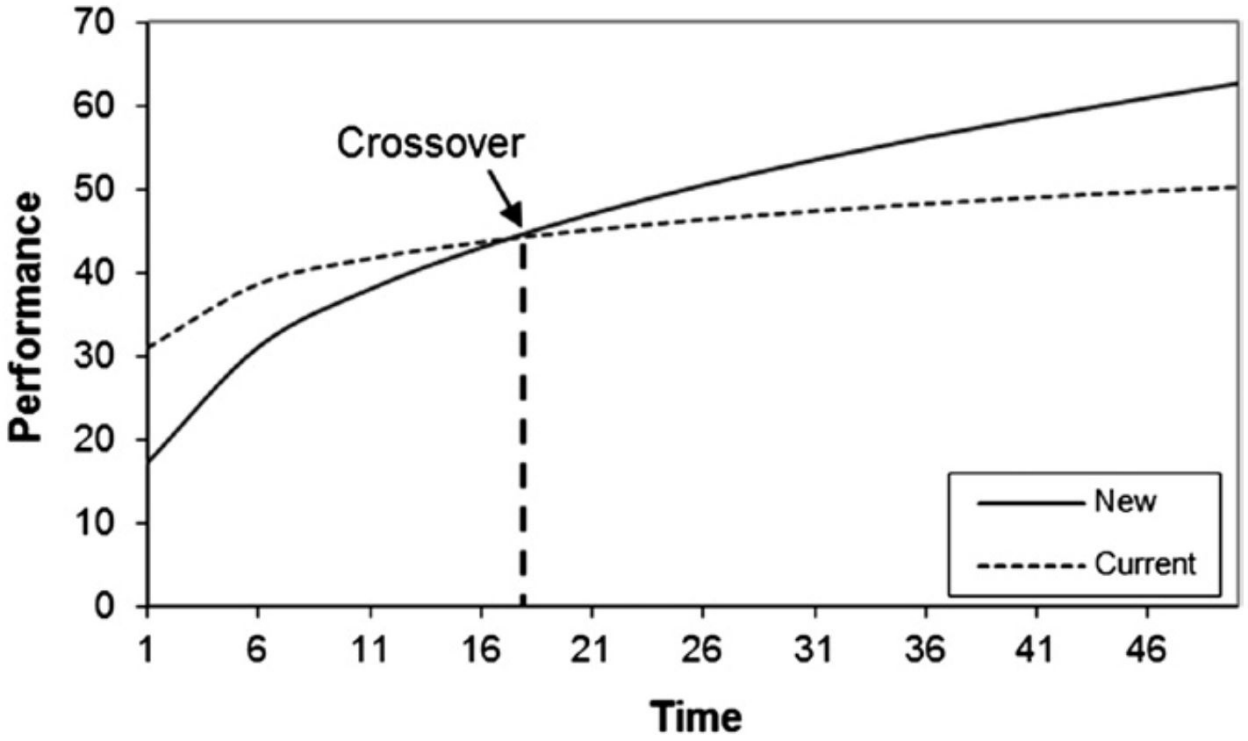
- This can be prevented either by between-subject design, or long enough training in within-subject design



Longitudinal studies

Sometimes, we are interested in learning or skill acquisition – *power law of learning*

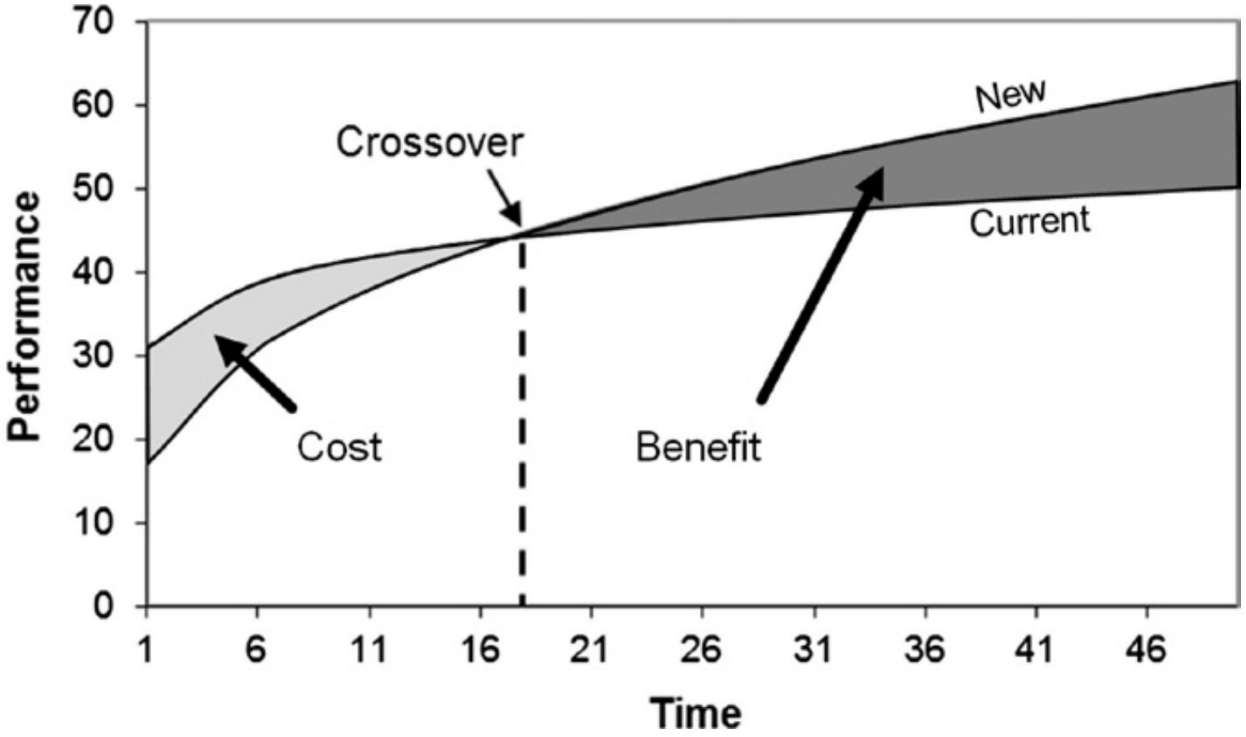
- "amount of practice" is an independent variable
- text entry, editing, pointing, selecting, searching, panning, zooming, rotating,...



Longitudinal studies

Sometimes, we are interested in learning or skill acquisition – *power law of learning*

- "amount of practice" is an independent variable
- text entry, editing, pointing, selecting, searching, panning, zooming, rotating,...



Running the experiment

pilot test (yes, one more pilot test) with one or two participants

1. Experimenter greets each participant
2. Introduces the experiment,
   Asks the participants to sign consent forms
3. Questionnaire is administered to gather demographic data and information on the participants' related experience
4. Apparatus is revealed, the task explained and demonstrated
5. Practice trials are allowed, as appropriate

Most interaction tasks, the participant is expected to proceed **quickly and accurately**

- quickly and accurately – are subject to interpretation, as the capabilities of participants

Bibliography

Lecture based on

MacKenzie, I. Scott. Human-computer interaction: An empirical research perspective. Newnes, 2012. (available online, google it)

Further reading

“Personal Dynamic Media” by A. Kay and A. Goldberg (1977).

“The Computer for the 21st Century” by M. Weiser (1991).