



DCGI

DEPARTMENT OF COMPUTER GRAPHICS AND INTERACTION
CZECH TECHNICAL UNIVERSITY IN PRAGUE



EMPIRICAL STUDIES POWER ANALYSIS

SAN 2023/24

HUMAN FACTORS

HUMAN FACTORS | VARIABILITY

- Humans are complicated – Computers are simple
- Age, 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

HUMAN FACTORS | TIME SCALE

- Workplace habits, groupware usage patterns, social networking, online dating, privacy, media spaces, design theory, ...
- Web navigation, user search strategies, collaborative computing, ubiquitous computing, social navigation, ...
- Selection techniques, force or auditory feedback, text entry, gestural input, ...

HUMAN FACTORS | TIME SCALE

- Workplace habits, groupware usage patterns, social networking spaces, design spaces, design
- Web navigation, collaborative social navigation
- Selection techniques, text entry, graphical

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 μ s	Organelle	

Newell 1999

HUMAN FACTORS | TIME SCALE

- Workplace habits, groupware usage patterns, social networking, online dating, privacy, media spaces, design theory, ...
- Web navigation, user search strategies, 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 μ s	Organelle	

Newell 1999

HUMAN FACTORS | TIME SCALE

Qualitative

Quantitative

- Workplace habits, groupware usage patterns, social networking, online dating, privacy, media spaces, design theory, ...
- Web navigation, user search strategies, 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 μ s	Organelle	

Newell 1999

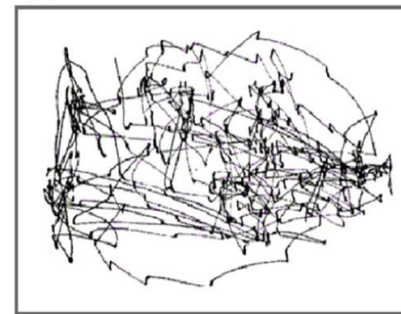
HUMAN FACTORS | SENSORS

- Vision
 - Intensity, Fixations, Saccades
- Hearing
 - Loudness, Pitch, Timbre
- Touch
 - Position, Texture, Temperature, Movement, Resistance

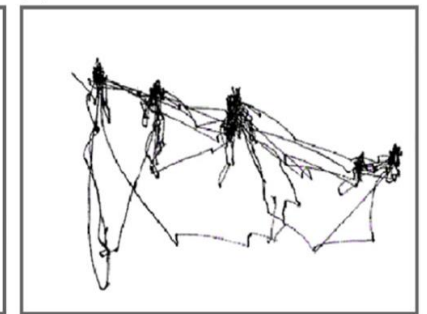
(a)



(b)



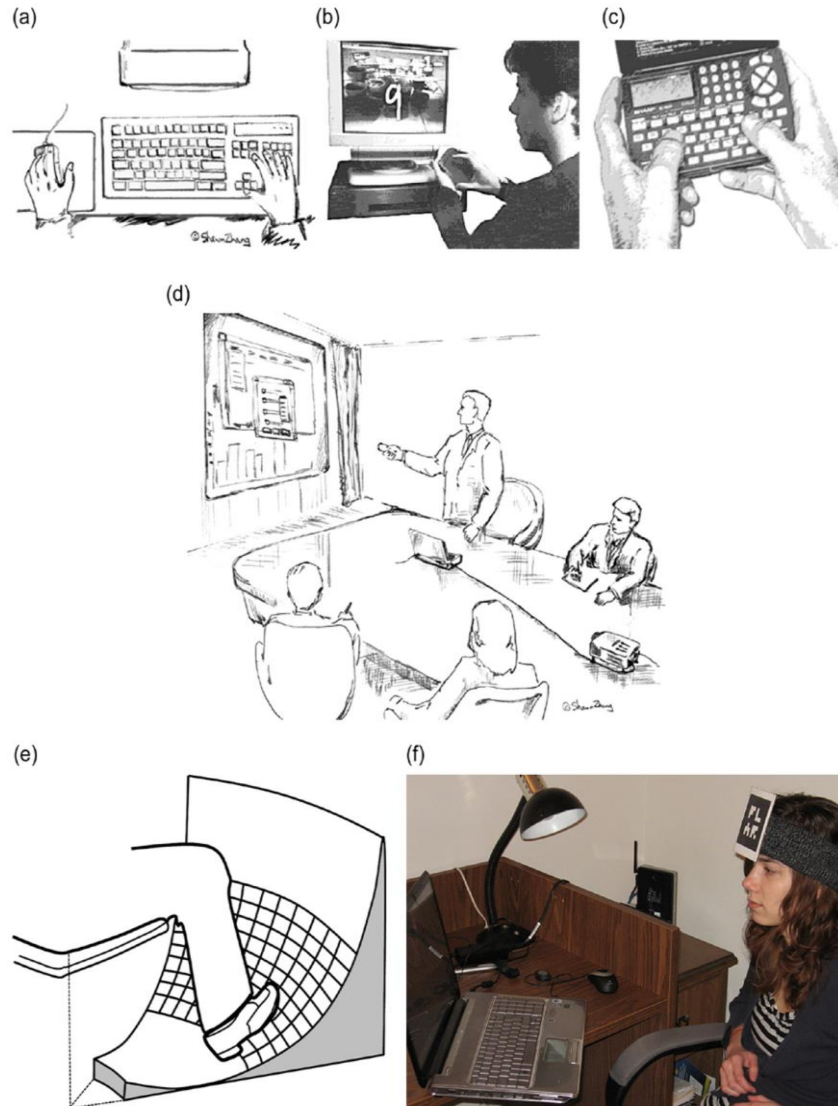
(c)



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

HUMAN FACTORS | RESPONDERS

- Limbs
- Voice
- Eyes
- Taste and smell

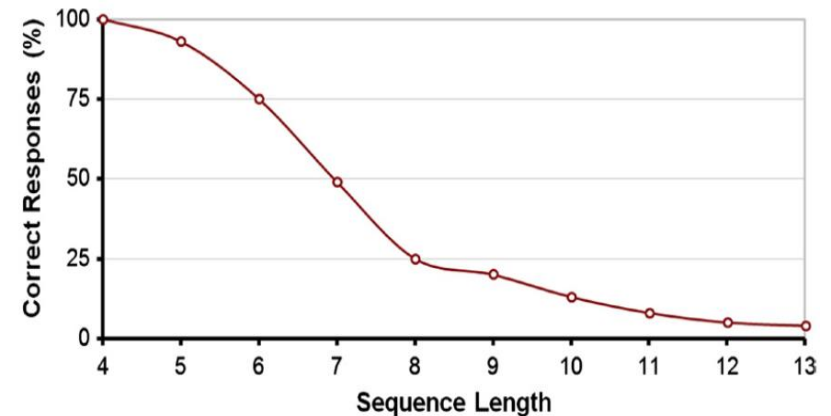


Use of the limbs in HCI: (a) Hands. (b) Fingers. (c) Thumbs. (d) Arms. (e) Feet. (f) Head.

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

HUMAN FACTORS | BRAIN

- Cognition
 - Thinking, reasoning, and deciding
- Memory
 - Long-term vs short-term (working)
- Language
 - Corpus, redundancy, entropy



```
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-----
```

HUMAN FACTORS | PERFORMANCE

- Reaction time
 - stimuli->response delay
- Time to make decision
 - logarithmic if there is a system
- Visual search
 - linear relation to number of items
- Skilled behavior
 - performance improves through training
- Attention
 - no cognitive action without attention
- Error
 - error is a discrete event in a task, or trial, where the outcome is incorrect

RESEARCH METHODS

RESEARCH METHODS

- Observation
- Experiment
- Correlation



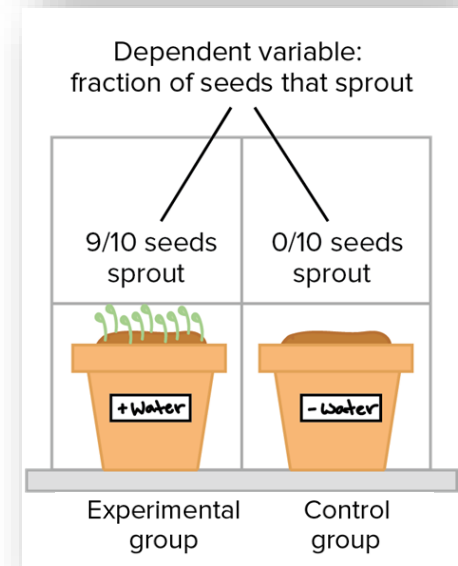
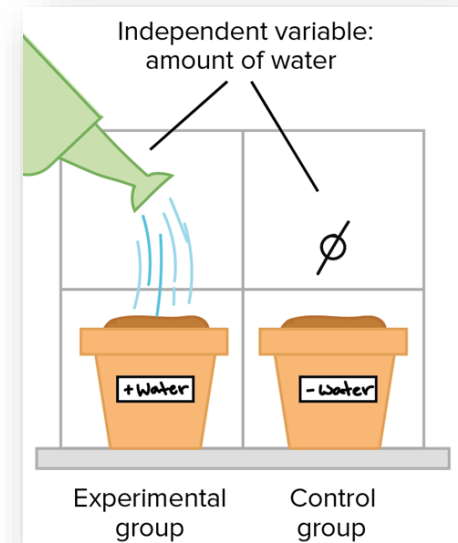
RESEARCH | OBSERVATION

- Interviews, field investigations, contextual inquiries, case studies, focus groups, ...
- Focus on thought, feeling, attitude, emotion, reaction, expression, sentiment, opinion, mood, manner, strategy, ...
- Qualitative rather than quantitative
- Achieves relevance while sacrificing precision



RESEARCH | EXPERIMENT

- Controlled experiments in laboratory settings
- Checking causality
 - manipulated (independent) variable => response (dependent) variable
 - systematically exposing participants to different configurations of the interface or interaction technique
- Measurement of responses
 - task completion time, number of errors, ...
- Allows conclusion to be drawn
 - hypothesis test



RESEARCH | CORRELATION

- Looking for relations between variables
- Quantification of variables is necessary
 - age, income, number of privacy settings
 - nominal-scale variables are categorized (e.g., personality type, gender)
- Data collected through a various methods
 - observation, interviews, on-line surveys, questionnaires, or measurement
- Balance between relevance and precision

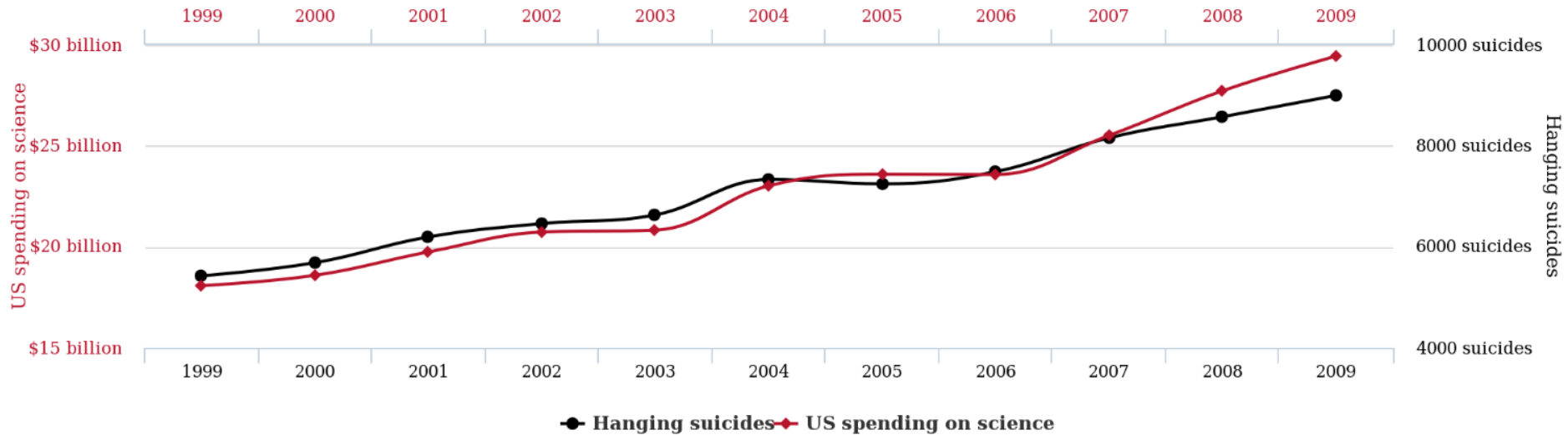
RESEARCH | CORRELATION

- Looking for relations between variables

US spending on science, space, and technology

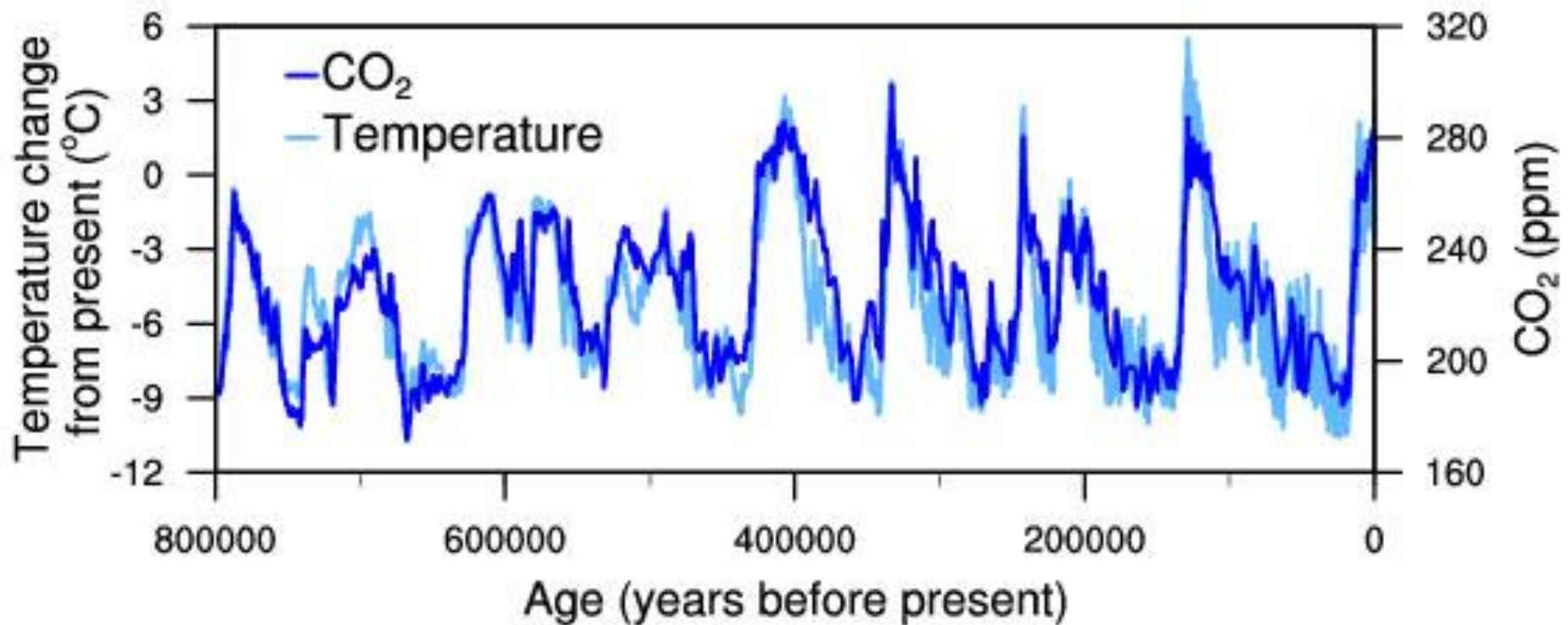
correlates with

Suicides by hanging, strangulation and suffocation



RESEARCH | CORRELATION

- Looking for relations between variables



NOAA, Jouzel 2007

MEASUREMENT

MEASUREMENT | SCALES

- Nominal, ordinal, interval, ratio
- Different sort of information
- Different analysis possible

MEASUREMENT | 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	ALBH	C	5
P13	M	ALBH		
P14	M	BLAH		
P15	F	BHAL		
P16	F	BLAH		
P18	M	BLAH		
P19	F	ALBH		
P20	M	AHBL		

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%		

MEASUREMENT | ORDINAL

- Order or ranking
- 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

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

MEASUREMENT | INTERVAL

- Equal distances between adjacent values
- There is no absolute zero
- Mean can be computed
- 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

Mackenzie 2013

MEASUREMENT | RATIO

- Ratio data have an absolute zero
- Time
 - completion time
- Count
 - normalization is recommended
- Errors normalized as “error rates (%)”
 - $\text{number of errors} / \text{number of trials} * 100$
 - $\text{number of incorrectly entered characters} / \text{total number of characters} \text{ times } 100$

RESEARCH QUESTION IN HCI

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

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

**Relevant, but
not testable!**

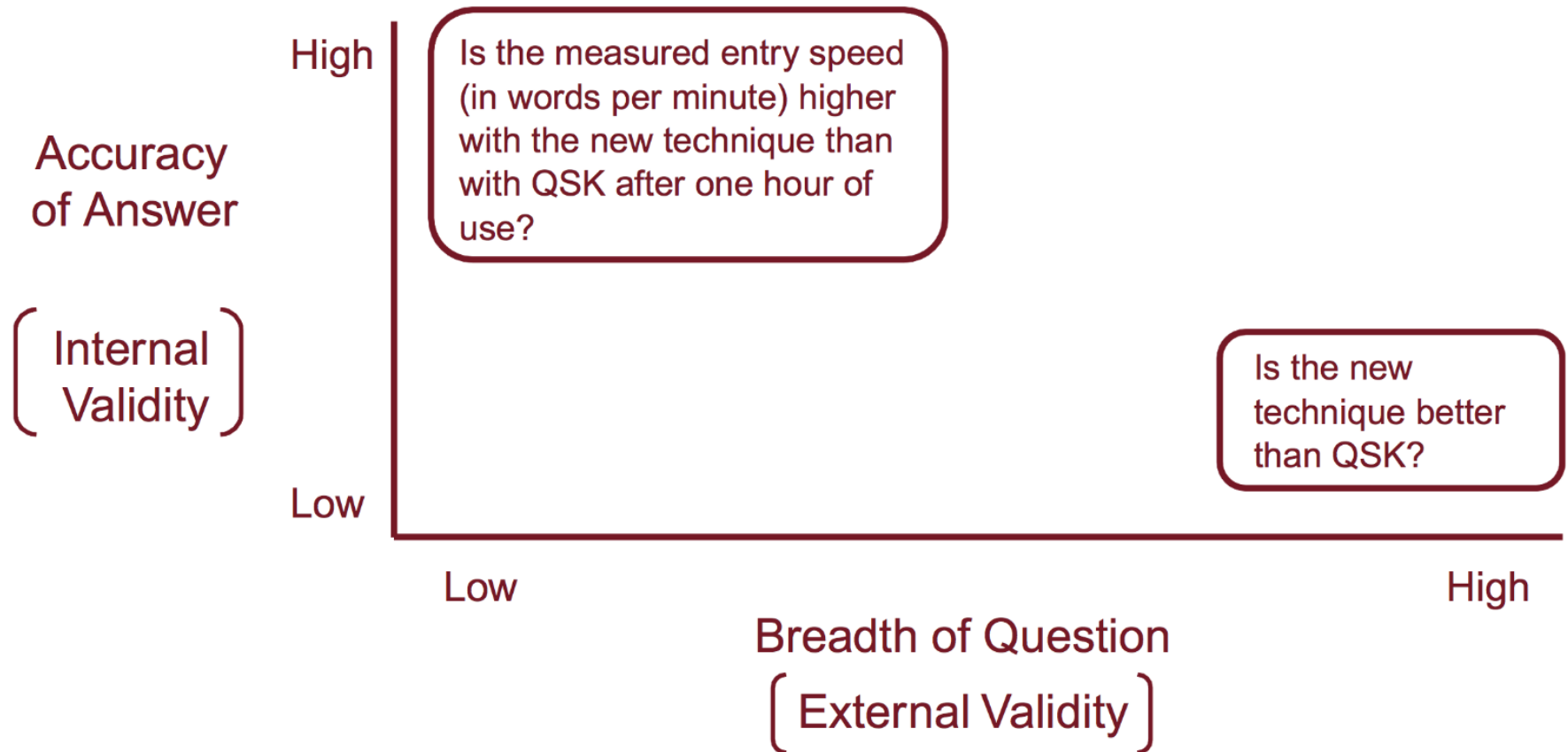
RESEARCH QUESTION

Example, questions about new input technique comparing to *Dvorak* 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



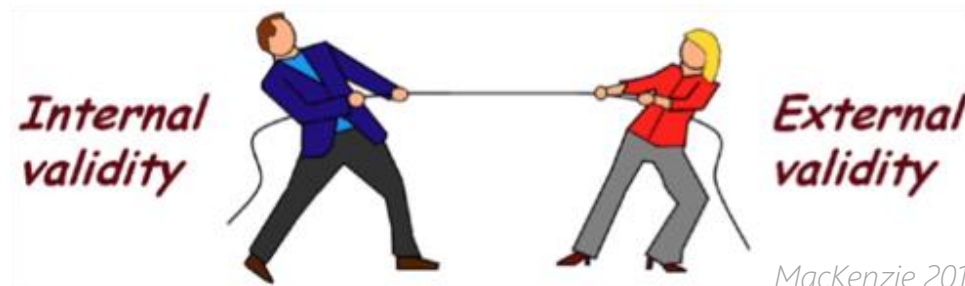
INTERNAL VS. EXTERNAL VALIDITY

- 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"



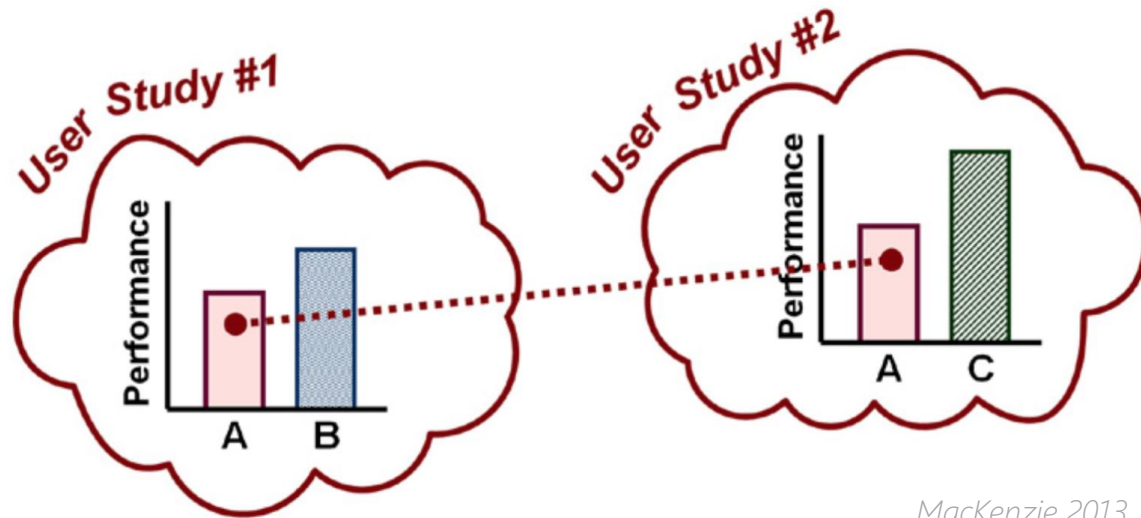
VARIABILITY AND CONFIDENCE

- People exhibit variability in their actions
- Variability person per person, but also person per task
- The result is always different!
- Variability strongly affects the confidence with which we can answer research questions

DESIGNING HCI EXPERIMENT

COMPARATIVE EVALUATION

- Evaluation on its own is questionable
- Baseline condition validates the methodology
- Testable research questions are crafted as comparisons



MacKenzie 2013

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

VARIABLES | INDEPENDENT

An independent variable (factor) is a characteristic that is manipulated or systematically controlled to evoke a change in a human response.

- Manipulated across multiple levels (at least 2)
- Independent of participant behavior
- Typically a nominal-scale attribute, often related to a property of an interface
 - device, entry method, feedback modality, selection technique, menu depth, button layout
 - unchangeable human characteristic (age, handedness, gender, expertise, ...)
 - environment characteristics (room lightning, noise, ...)

VARIABLES | DEPENDENT

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, ...
- Dependent on the human behavior
- Any observable, measurable aspect of human behavior is a potential dependent variable
 - all dependent variables must be clearly defined to ensure the research can be replicated

VARIABLES | OTHER

■ Control variables

- influence a dependent variable but are not under investigation => we try to make them constant
- lighting, temperature, noise, display size, mouse shape, keyboard angle, chair height, participant characteristic

■ Random variables

- increase variability of measured behavior => results are less accurate
- typically characteristics of the participants: biometrics, 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.

MacKenzie 2013

VARIABLES | OTHER

- 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 (difference in conditions), ...

VARIABLES | 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

TASK & PROCEDURE

- Procedure should contain all combinations of independent variable and their values
- Task is representative and discriminates
- Besides tasks the procedure contains instruction and training

PARTICIPANTS

- Select participants from the same population to whom to results apply
- Use sufficient number of participants
 - a priori power analysis
 - check similar research studies
- Increasing the number of participants increases the likelihood of achieving statistically significant results
 - Large number of participants: statistically significant results for a difference of no practical significance

PARTICIPANTS | WITHIN/BETWEEN S.

WITHIN-SUBJECT

- repeated measures
- less participants
- variance low
- interference between test cond.
 - learning effect
 - fatigue effect

BETWEEN-SUBJECT

- separate groups
- more participants
- balancing needed
- no interference between test cond.

PARTICIPANTS | CONTERBALANCING

- 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**
- $n \times n$ 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

PARTICIPANTS | CONTERBALANCING

- **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

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

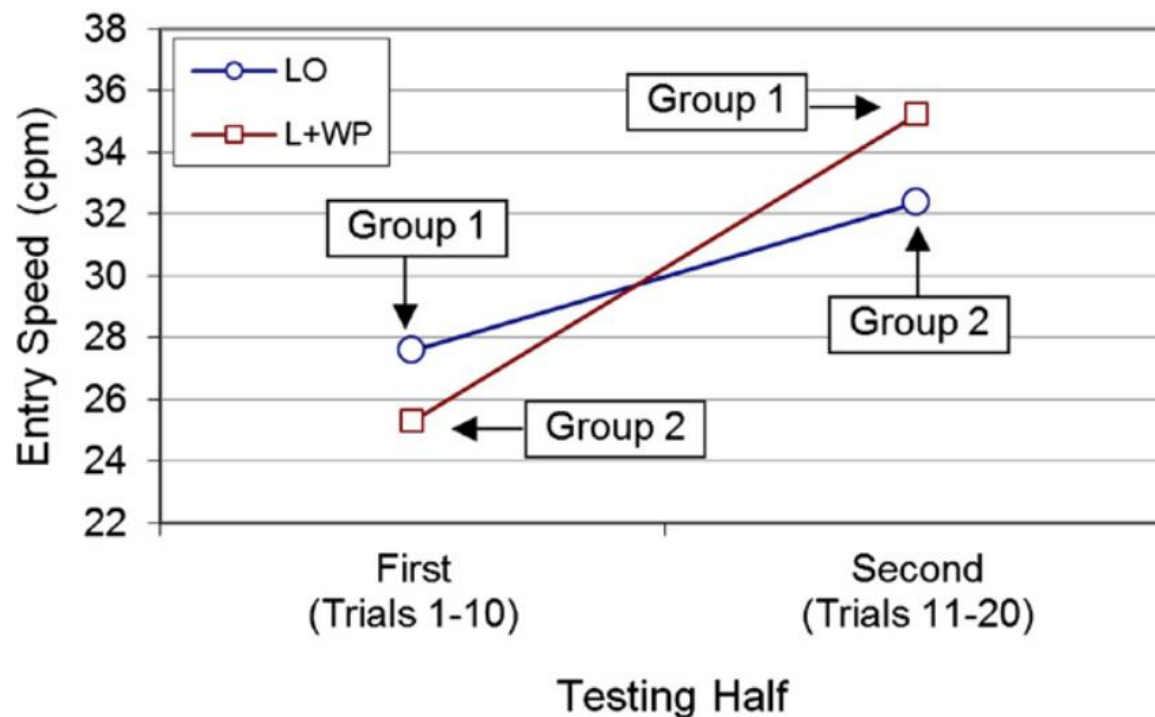
(b)

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 .

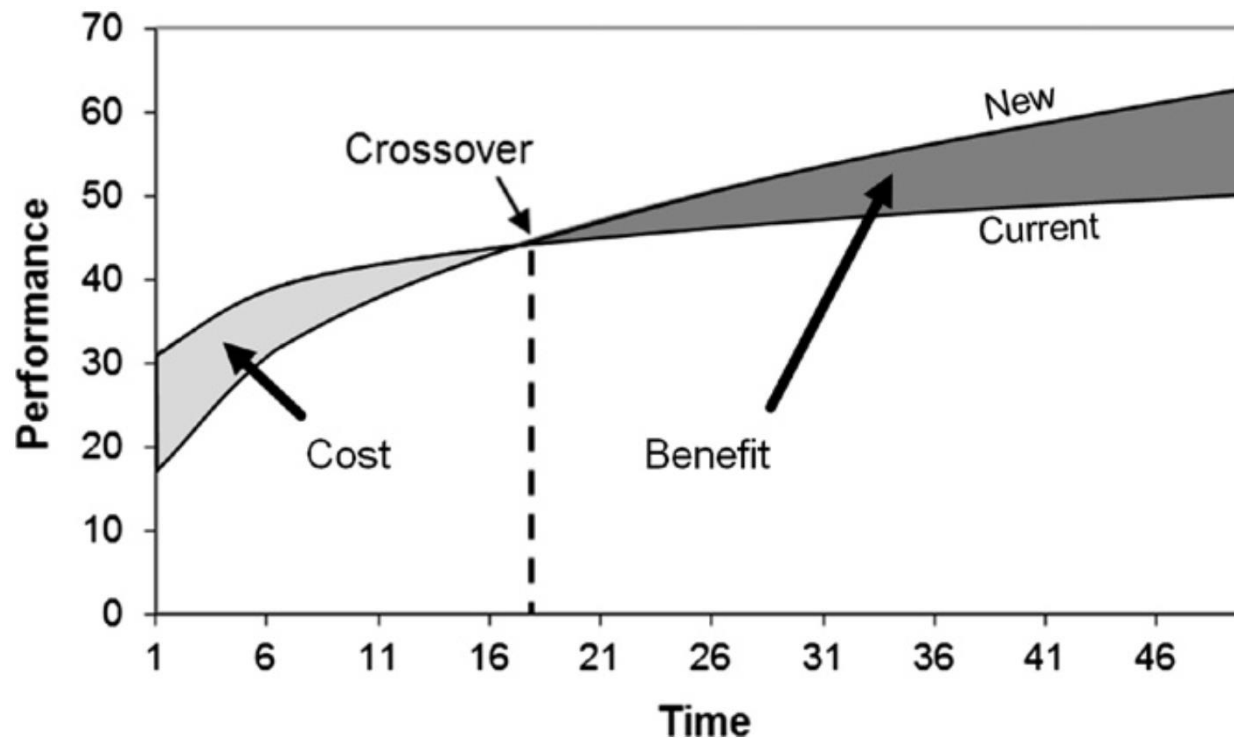
ASYMMETRIC SKILL TRANSFER

- There are occasions where different learning effects appear for one order (e.g., A→B) compared to another (e.g., B→A)
 - group effect = different amount of improvement depending on the order of testing



LONGITUDINAL STUDIES

- Sometimes, we are interested in learning or skill acquisition – power law of learning
- “amount of practice” is an independent variable



POWER ANALYSIS

ERRORS IN EXPERIMENTS

- Type I error (False positive, α error)
 - H_0 is rejected, when in reality H_1 is not correct
- Type II error (False negative, β error)
 - H_0 is not rejected (H_1 is not accepted), when in reality H_1 is correct

	H0 not rejected	H1 accepted
H0 is truth	Correct	Type I error
H1 is truth	Type II error	Correct

SOURCES OF ERRORS

- 1. Usability properties identification
- 2. Prototype creation
- 3. Experiment design
- 4. Participants recruitment
- 5. Test execution and data collection
- 6. Data analysis
- 7. Conclusions and recommendations statement

SOURCES OF ERRORS | CONT.

■ 3. Experiment design

- poor choice of stimuli
- wrong choice of task
 - unaware of the task
 - poor design in terms of instructions
- accidental errors
- insignificant stimuli
 - large spread of stimuli
 - shift of modality



■ 6. Data analysis

- analysis of influence of test conditions on the data measured
- evaluator bias => analysis performed by more evaluators

DATA ANALYSIS | OUTLIERS

- Outliers are always there
 - but more often for “long tail” distributions
- Outliers elimination
 - selection bias => “data fishing”
 - before looking at the data measured (step 6)
 - better: before test execution (step 5)
 - perform qualitative evaluation of outliers behavior

	method A					method B				
min	26	24	22	17	15	10	9	8	7	6
max	94	98	75	82	72	41	39	31	29	27

SAN 2018 experiment

POWER ANALYSIS

- Power of a test = $(1 - \beta)$
 - probability that the test correctly rejects H_0

$$\text{power} = \mathbb{P}(\text{reject } H_0 | H_1 \text{ is true})$$

- Depends on
 - significance level α (Type I error probability)
 - sample size n
 - effect size d (min. degree of violation of H_0)
 - specify on a priori grounds

$$\text{t test: } \text{Cohen's } d = \frac{\mu_1 - \mu_2}{\sigma}$$

POWER ANALYSIS | SIZE d

- t tests

- Cohen's suggestion:
0.2, 0.5, 0.8

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- ANOVA

- Cohen's suggestion:
0.1, 0.25, 0.4

$$f = \sqrt{\frac{\sum_{i=1}^k p_i * (\mu_i - \mu)^2}{\sigma^2}}$$

$$p_i = n_i/N$$

n_i = number of observations in group i

μ = grand mean

- Chi-square test

- Cohen's suggestion:
0.1, 0.3, 0.5

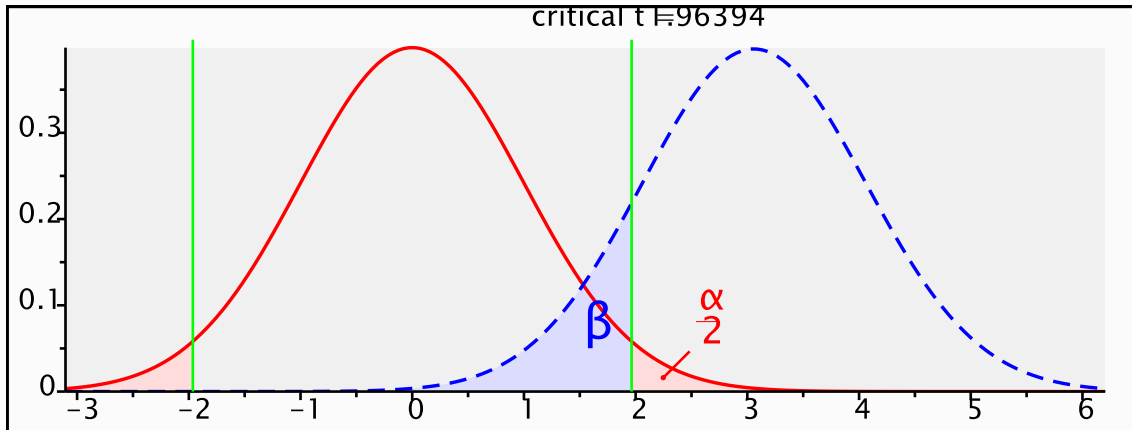
$$w = \sqrt{\sum_{i=1}^m \frac{(p0_i - p1_i)^2}{p0_i}}$$

$p0_i$ = cell probability in i^{th} cell under H_0

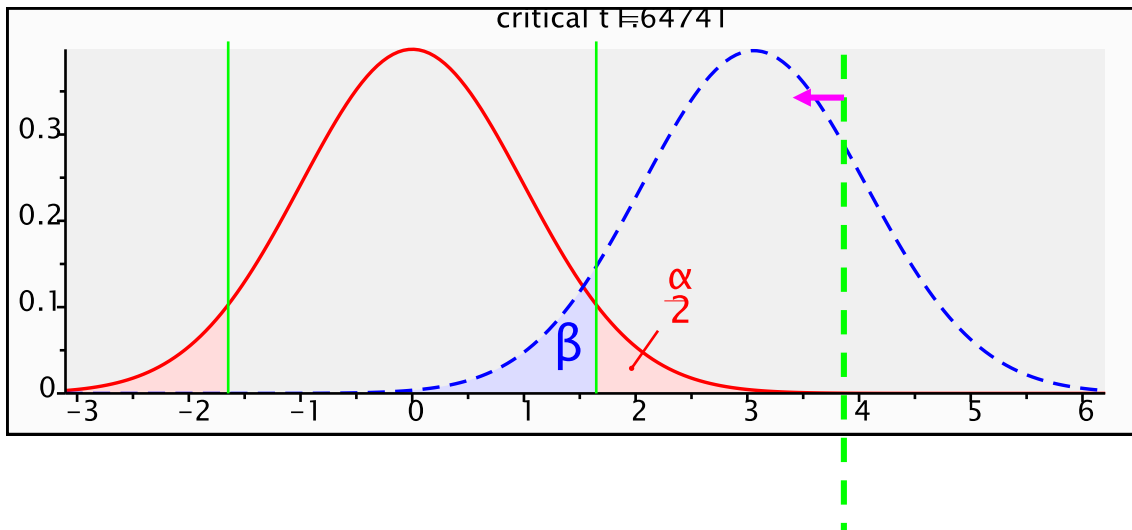
$p1_i$ = cell probability in i^{th} cell under H_1

POWER ANALYSIS | DEPENDENCE

t test (difference between two independent means)



$$\alpha = 0.05$$
$$\beta = 0.14$$



$$\alpha = 0.1$$
$$\beta = 0.08$$

POWER ANALYSIS | TYPES

- A priori
 - controlling power level before conducting test
 - computing sample size n
 - function of required power level, specified α , d
- Post hoc
 - after a test was conducted
 - Does the test had fair chance to reject incorrect H_0 ?
 - computing the power level
- Compromise
 - fixed ratio between α and β
- Sensitivity
 - estimating/checking the size of an effect d

POWER ANALYSIS | DISCOVERY

- How many users do we need for discovering 95% of (**ALL**) problems?
- Golden rule of usability testing: Five users is enough to observe **all relevant** problems with **very high** probability.
- To detect X % of problems that affects Y % of users.
- To have a X % chance of detecting ...

$$n = \frac{\ln(1 - X)}{\ln(1 - Y)}$$

$$n = 5$$

very high = 95 %

all relevant = 50 %

POWER ANALYSIS | COMPARING

- Determining n for comparing two means
 - within-subject

$$n = \frac{(t_{\alpha} + t_{\beta})^2 s^2}{d^2}$$

t_{α} = critical value for Confidence level

t_{β} = critical value for Power

s^2 = the variance (estimate of SD^2)

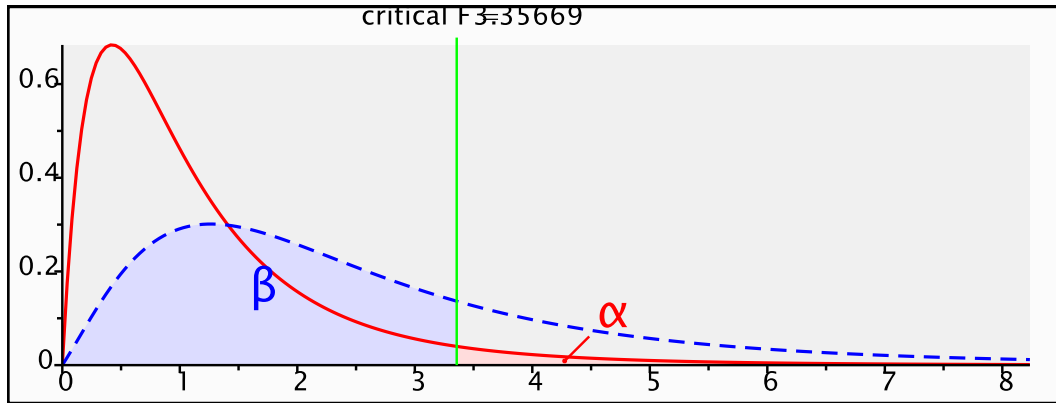
d^2 = the square of critical difference

- between subject

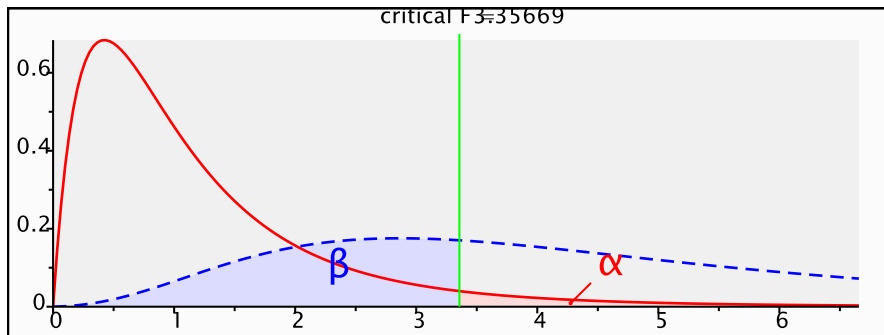
$$n = \frac{2(t_{\alpha} + t_{\beta})^2 s^2}{d^2}$$

POWER ANALYSIS | COMPARING

F test (MANOVA: Repeated measures, within factors)



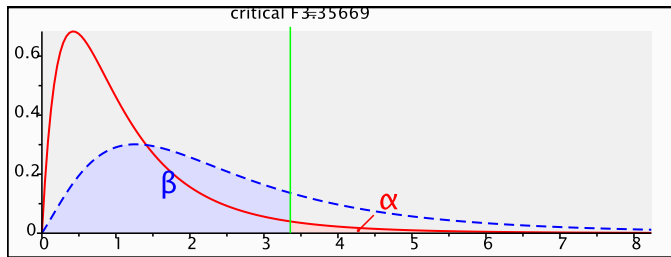
$\alpha = 0.05$
 $\beta = 0.73$
 $f = 0.25$ (medium)
 $n = 16$



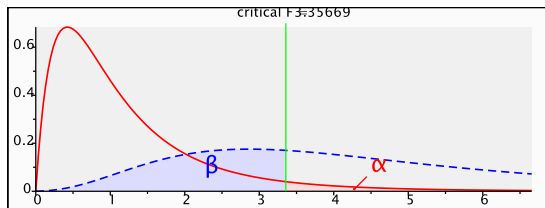
$\alpha = 0.05$
 $\beta = 0.37$
 $f = 0.4$ (large)
 $n = 16$

POWER ANALYSIS | COMPARING

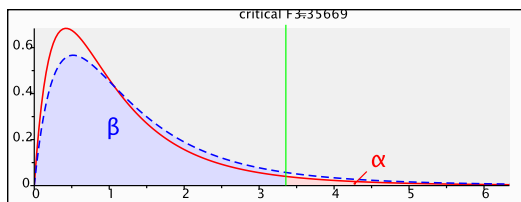
F test (MANOVA: Repeated measures, within factors)



$\alpha = 0.05$
 $\beta = 0.73$ for $\beta = 0.2, n = 44$
 $f = 0.25$ (medium)
 $n = 16$



$\alpha = 0.05$
 $\beta = 0.37$ for $\beta = 0.2, n = 22$
 $f = 0.4$ (large)
 $n = 16$



$\alpha = 0.05$
 $\beta = 0.92$ for $\beta = 0.2, n = 244$
 $f = 0.1$ (small)
 $n = 16$

EXPERIMENT RESULTS

F test (MANOVA: Repeated measures, within factors)

Keyboard type means:

A=41.86400

B=14.40800

Group means:

AB=29.92800

BA=26.34400

```
=====
```

Effect	df	SS	MS	F	p
Group	1	1605.632	1605.632	3.020	0.08865
Participant (Group)	48	25519.320	531.653		
Keyboard type	1	94228.992	94228.992	341.435	0.00000
Keyboard type x Group	1	1083.392	1083.392	3.926	0.05330
Keyboard type_x_P (Group)	48	13247.016	275.979		
Trails	4	8265.372	2066.343	107.509	0.00000
Trails_x_Group	4	38.148	9.537	0.496	0.73855
Trails_x_P (Group)	192	3690.280	19.220		

```
=====
```

SAN 2018 experiment

THANK YOU FOR ATTENTION



DCGI

DEPARTMENT OF COMPUTER GRAPHICS AND INTERACTION
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Zdeněk Míkovec
xmikovec@fel.cvut.cz

REFERENCES

- 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).

REFERENCES

- Power Analysis in R: <http://www.statmethods.net/stats/power.html>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, ISBN 978-0805802832, Routledge.
- Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Elsevier.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). *Statistical power analyses using G* Power 3.1. 9: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences*. *Behavior Research Methods*, 41(4), 1149-1160.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). *Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses*. *Behavior research methods*, 41(4), 1149-1160.
- Kuniavsky, M. (2012). *Observing the user experience: a practitioner's guide to user research*. Morgan Kaufmann.
- MacKenzie, I. S. (2012). *Human-computer interaction: An empirical research perspective*. Newnes.