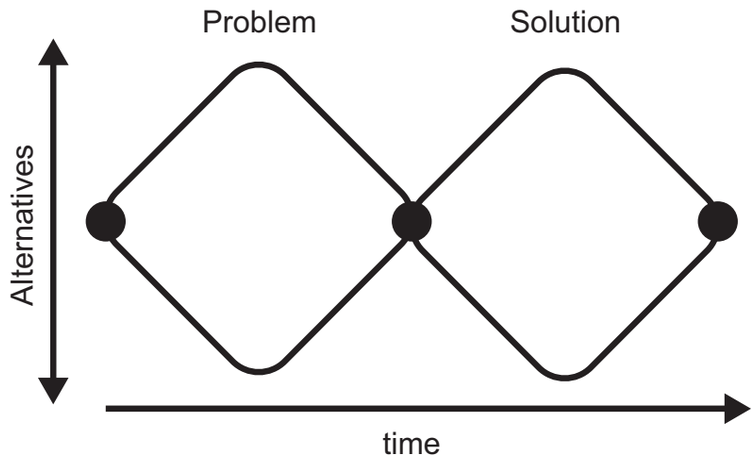
Evaluation and Testing

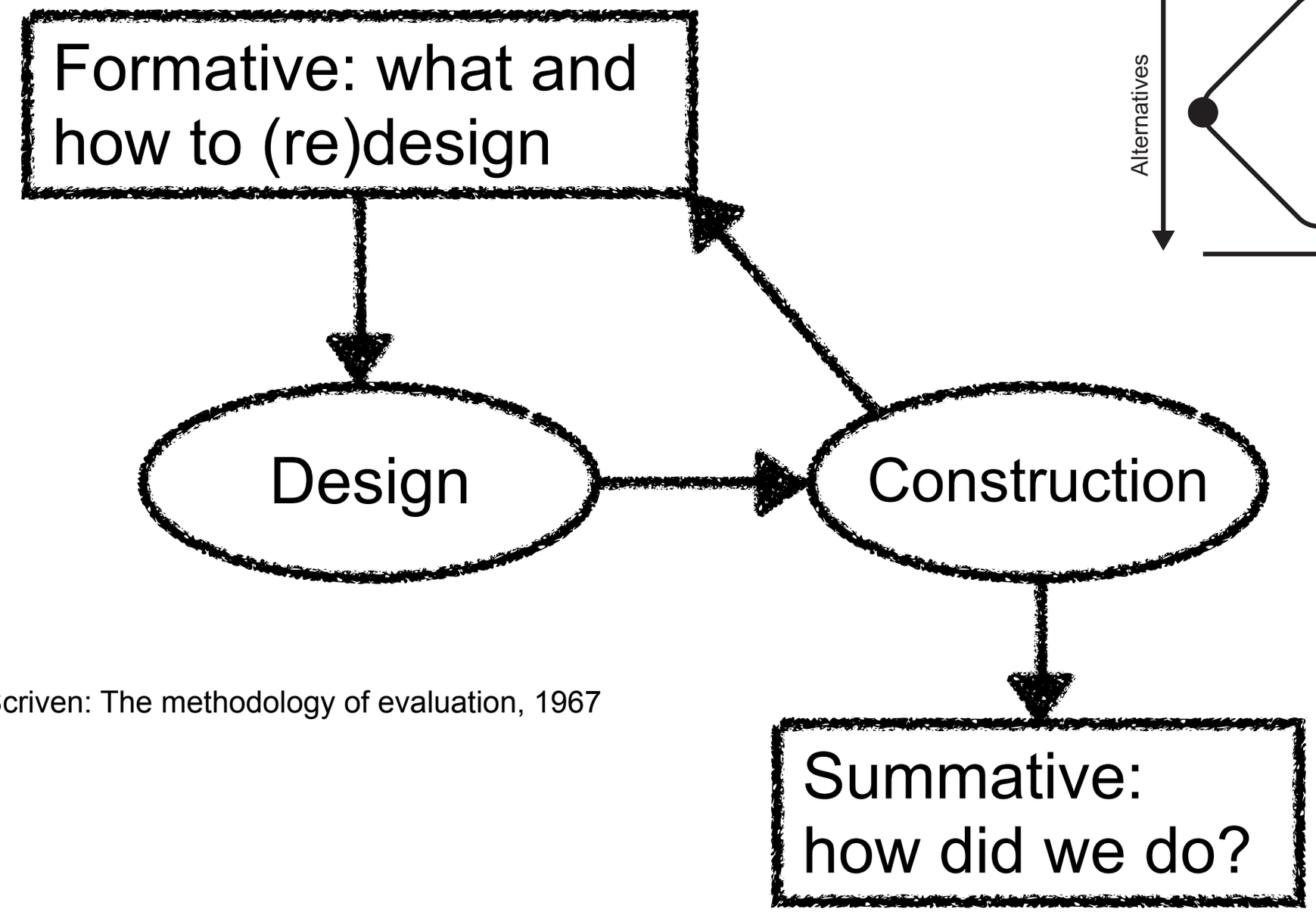
Andreas Butz, LMU Media Informatics
butz@lmu.de
slides partially taken from MMI class

The user as the ultima ratio...

Donald Norman

Formative vs. Summative Evaluation





- M. Scriven: The methodology of evaluation, 1967

Qualitative vs. Quantitative Evaluation

http://www.scope-mr.ch/de/dienstleistungen/methoden/

http://www.scope-mr.ch/de/dienstleistungen/methoden/

http://blog.efpsa.org/wp-content/uploads/2012/05/yin_yang.png

Analytic vs. Empirical Evaluation

Scriven, 1967: “If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you may just study the kind and speed of the cuts it makes in the hands of a good axeman.”

Empirical and Analytic Methods are Complementary (not complimentary ;-)

- Empirical evaluation produces facts which need to be interpreted
  - If the axe does not cut well, what do we have to change?
  - Analytic evaluation identifies the crucial characteristics

- Analytical evaluation produces facts which need to be interpreted
  - Why does the axe have a special-shaped handle?
  - Empirical evaluation helps to understand the context for object properties

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough GOMS + KLM	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough GOMS + KLM	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

Types of Analytical Evaluation

- Inspection-based evaluation
  - Expert review
  - Heuristic evaluation
  - Cognitive walkthrough
- Model-based evaluation
  - Evaluation according to models of how interaction works
- Different results
  - Qualitative assessment
  - Quantitative assessment

Cognitive Walkthrough

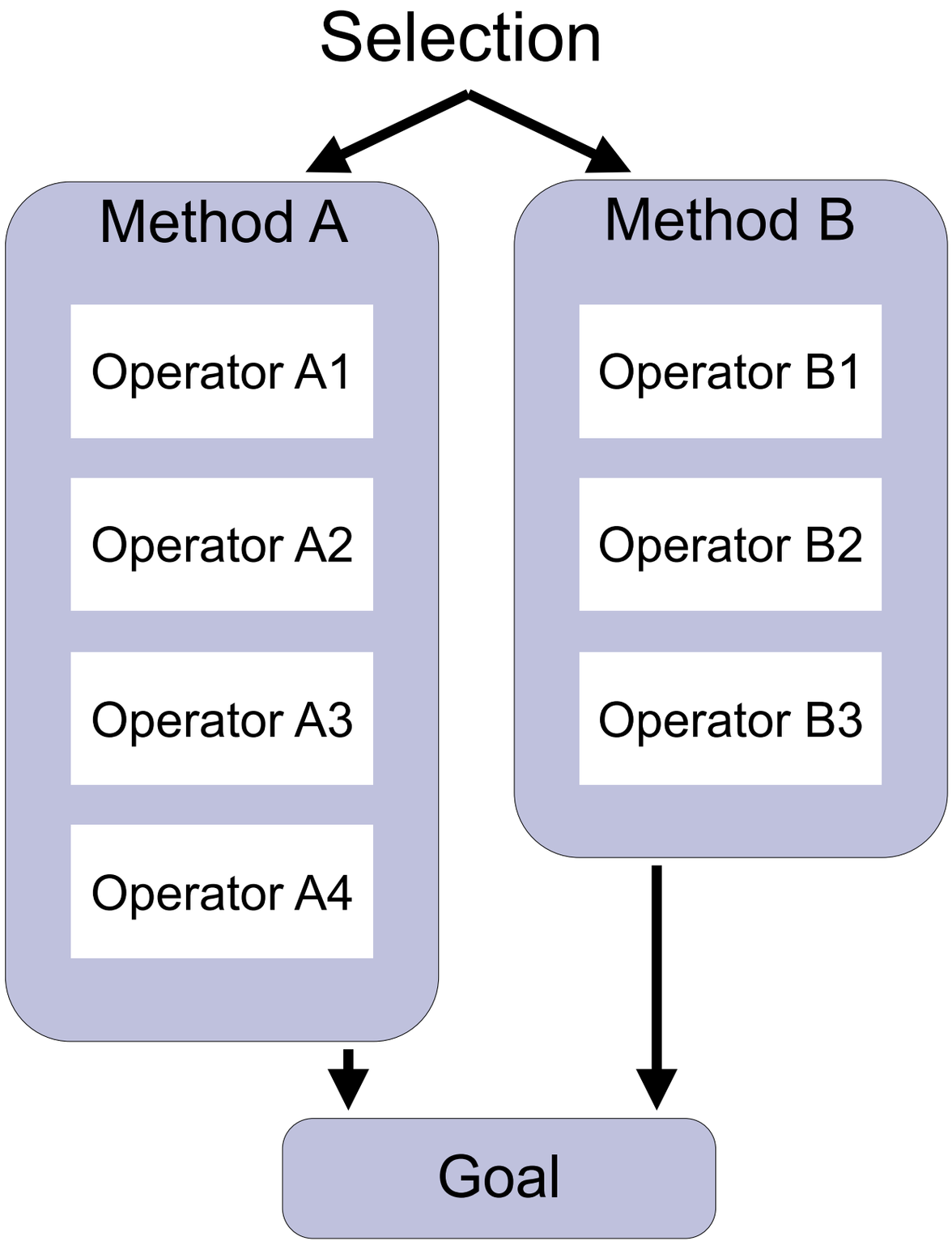
…Step by step…

…along well-defined tasks…

1. Is the **correct action** for executing the next step always clearly defined? Does the user know what to do next?
2. Is the correct action clearly **recognizable**? Does the user actually find it?
3. Does the user receive a sufficient **feedback** after executing the action, such that he can determine whether the action was executed successfully?

Goals, Operators, Methods & Selection Rules (**GOMS**)

- **S**election rules
- **M**ethods
- **O**perators
- **G**oals



Keystroke Level Model **(KLM)**

- Used times in experimental average:

- **K** (Keystroke): Pressing a key: $\mathbf{t_K = 0.28s}$.
- **P** (Pointing): Pointing to a position on screen: $\mathbf{t_P = 1.1s}$
- **H** (Homing): Switch between keyboard and mouse: $\mathbf{t_H = 0.4s}$
- **M** (Mental preparation): Mental preparation of successive operation: $\mathbf{t_M = 1.35s}$
- **R(t)** (Response time): Response time of the systems (within **t** seconds, system-dependent).

KLM example

1. point to file icon **P**
2. press and hold mouse button **B**
3. drag file icon to trash can icon **P**
4. release mouse button **B**
5. point to original window **P**

Total time = 3P + 2B = 3*1.1 + 2*.1 = 3.5 sec

ftp://www.eecs.umich.edu/people/kieras/GOMS/KLM.pdf

KLM example 2

- Which of the methods M1 or M2 is faster?

- **M1**: Switch to mouse, move mouse pointer to file icon, clicking the icon, dragging to trash icon and release, switch to keyboard

- **M2**: Switch to mouse, selecting the icon, switch to keyboard, press ‘delete’

- $\mathbf{t_{M1}} = t_H + t_P + t_K + t_P + t_H = 0.4 + 1.1 + 0.28 + 1.1 + 0.4 =$ **3.28s**
- $\mathbf{t_{M2}} = t_H + t_P + t_H + t_K = 0.4 + 1.1 + 0.4 + 0.28 =$ **2.18s**

KLM table

- **K** - Keystroke (.12 - 1.2 sec; .28 recommended for most users).
  - Expert typist (90 wpm): .12 sec
  - Average skilled typist (55 wpm): .20 sec
  - Average nonsecretarial typist (40 wpm): .28 sec
  - Worst typist (unfamiliar with keyboard): 1.2 sec
- **T(n)** - Type a sequence of n characters on a keyboard (n * K sec).
- **P** - Point with mouse to a target on the display (1.1 sec).
  - The actual time required can be determined from Fitts' law.
  - For typical situations, it ranges from .8 to 1.5 sec, with an average of 1.1 sec.
- **B** - Press or release mouse button (.1 sec).
- **BB** - Click and release mouse button (.2 sec).
- **H** - Home hands to keyboard or mouse (.4 sec).

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough GOMS + KLM	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

10 Usability Heuristics


Jakob Nielsen

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose, and recover from errors
- Help and documentation

Detailed Checklist Example

Usability Techniques
Heuristic Evaluation - A System Checklist

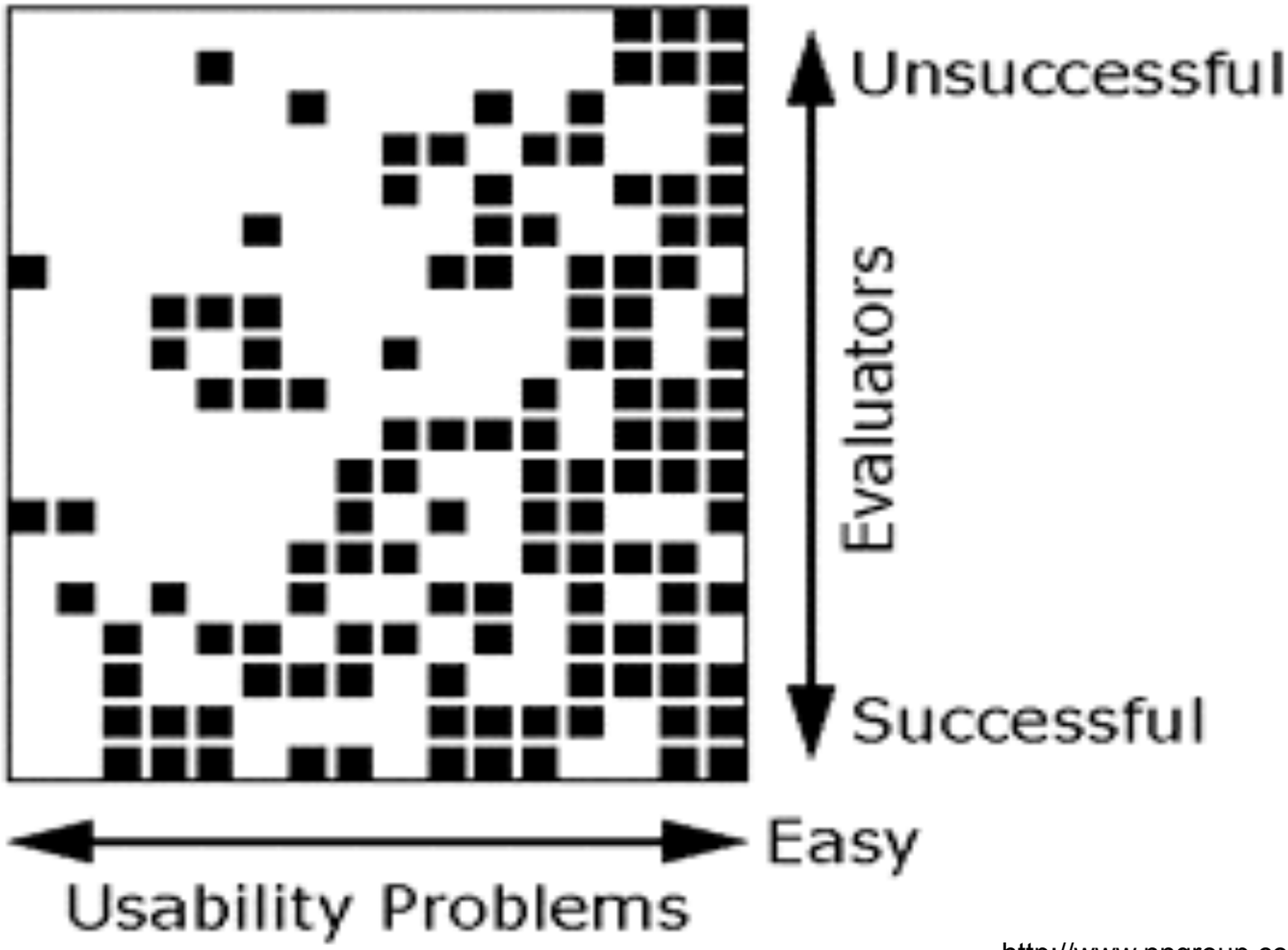
By Deniese Pierotti, Xerox Corporation

Heuristic Evaluation - A System Checklist http://www.stcsig.org/usability/topics/articles/he-checklist.html

1. Visibility of System Status

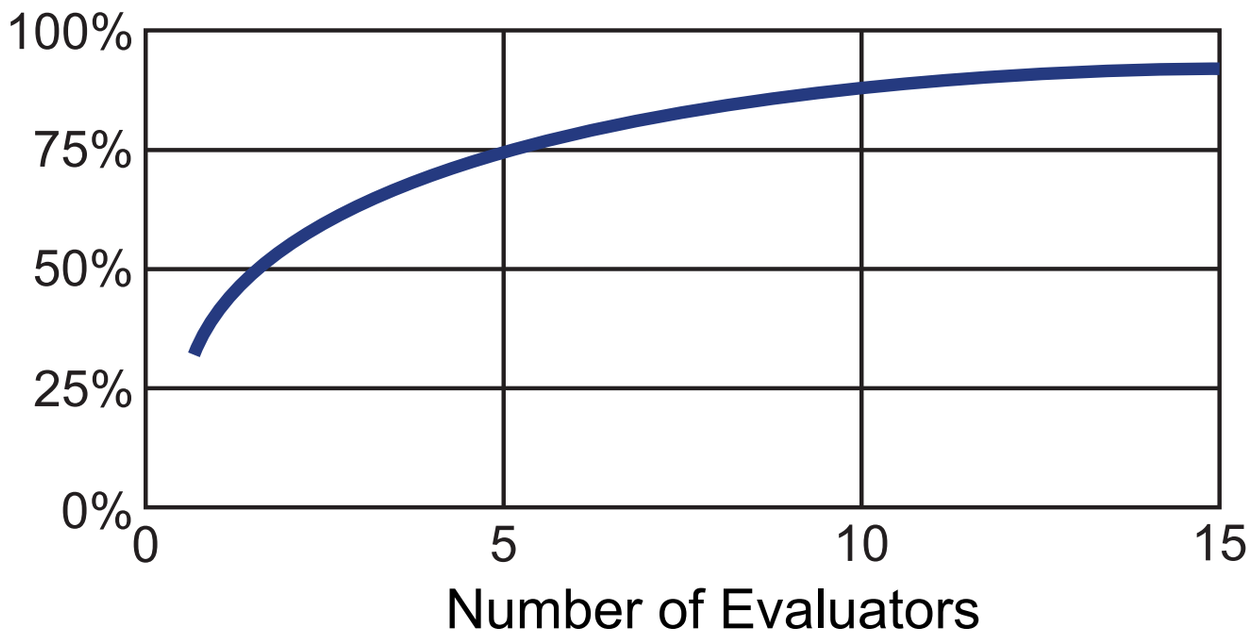
The system should always keep user informed about what is going on, through appropriate feedback within reasonable time.

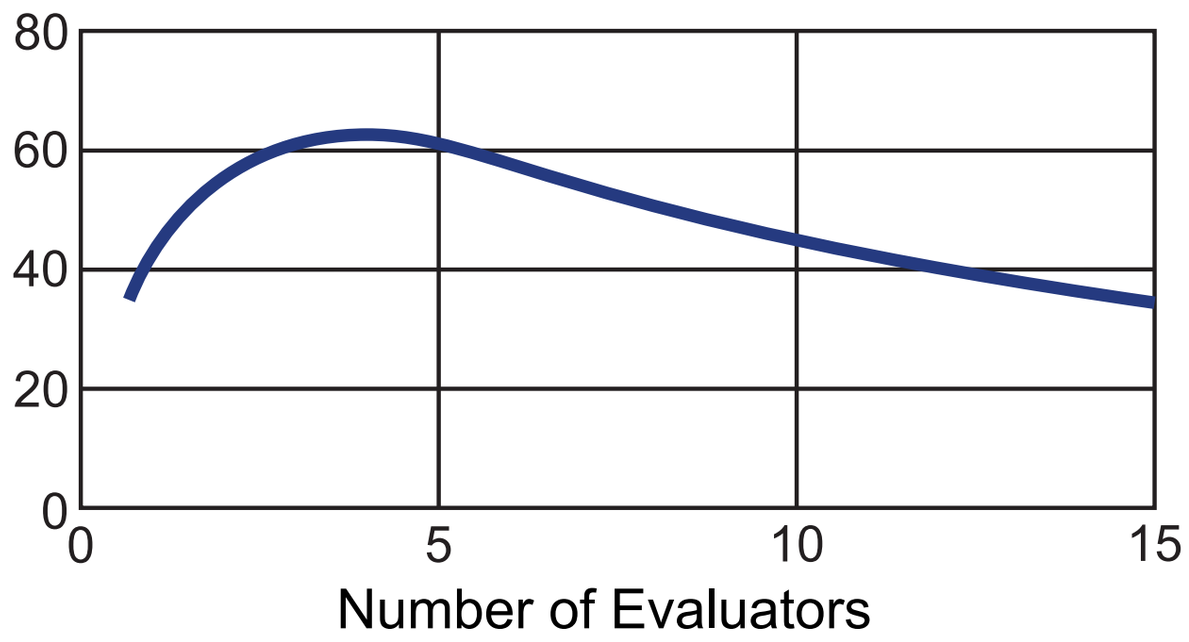
#	Review Checklist	Yes No N/A	Comments
1.1	Does every display begin with a title or header that describes screen contents?	O O O	
1.2	Is there a consistent icon design scheme and stylistic treatment across the system?	O O O	
1.3	Is a single, selected icon clearly visible when surrounded by unselected icons?	O O O	
1.4	Do menu instructions, prompts, and error messages appear in the same place(s) on each menu?	O O O	
1.5	In multipage data entry screens, is each page labeled to show its relation to others?	O O O	
1.6	If overtype and insert mode are both available, is there a visible indication of which one the user is in?	O O O	
1.7	If pop-up windows are used to display error messages, do they allow the user to see the field in error?	O O O	
1.8	Is there some form of system feedback for every operator action?	O O O	
1.9	After the user completes an action (or group of actions), does the feedback indicate that the next group of actions can be started?	O O O	
1.10	Is there visual feedback in menus or dialog boxes about which choices are selectable?	O O O	
1.11	Is there visual feedback in menus or dialog boxes about which choice the cursor is on now?	O O O	



Jakob Nielsen

http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/





Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough GOMS + KLM	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

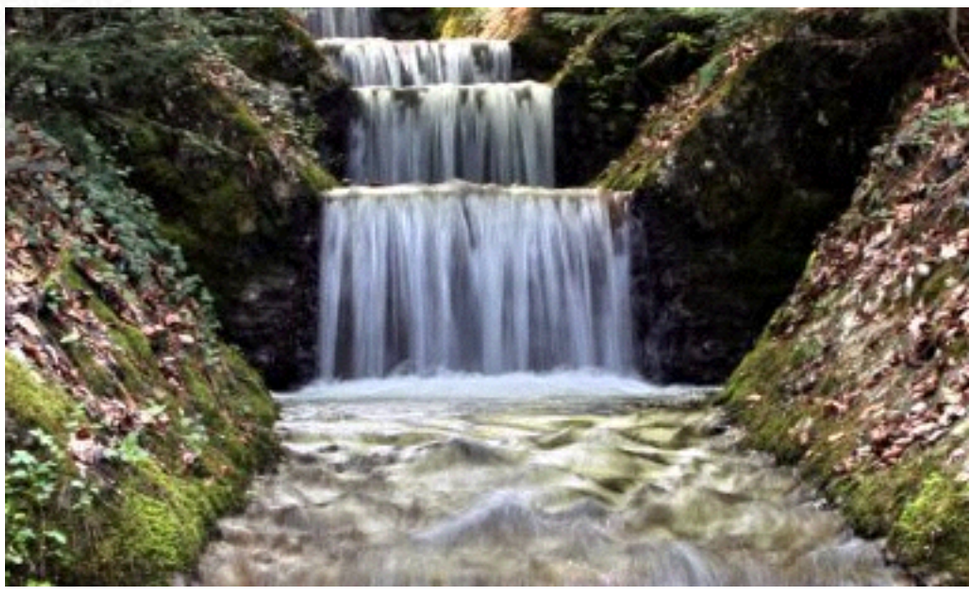
Quality Properties of Empirical Methods

- Objectivity
- Reproducability
- Validity
  - internal
  - external
- Relevance

http://www.schwimmvereinapolda.de/images/Webelemente/Stoppuhr.jpg

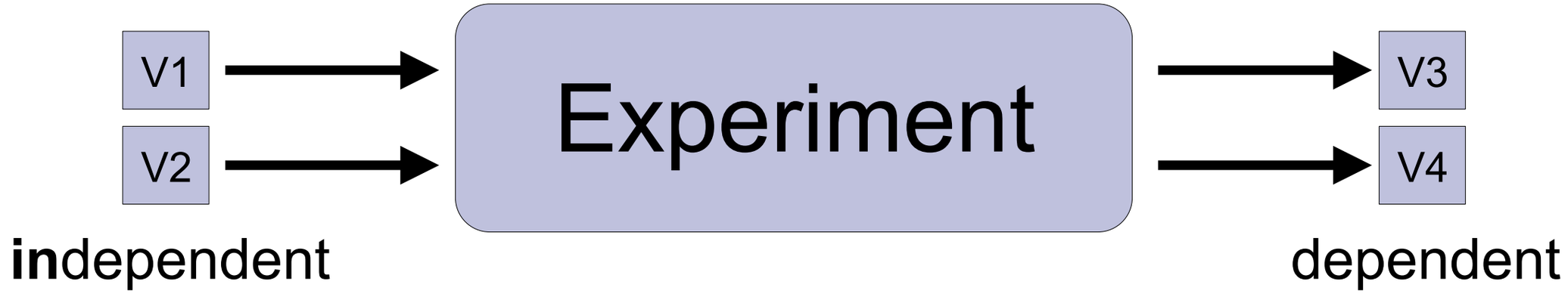
http://www.bgr.bund.de/DE/Themen/Endlagerung/Bilder/end_nfpro_hyperf_g.jpg?__blob=normal&v=2

http://wl15www815.webland.ch/travelinfos/images/mensch/gehirn4.jpg



http://bilder.n3po.com/cache/Photos/Bach-Fliessend-Bergab_w475_h230_cw475_ch230_thumb.jpg

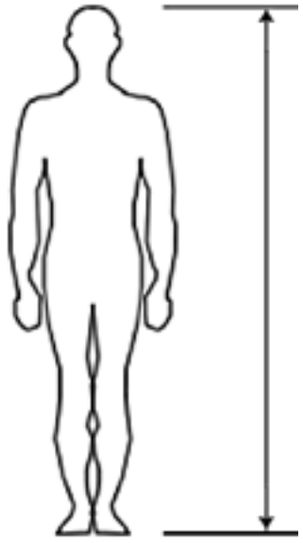
Variables and Values



- Nominal
- Ordinal
- Cardinal

http://www.gebr-clasen.de/-g400/weltkarte_KB.png

http://www.zukunftskinder.org/wp-content/uploads/2013/01/großfamilie.jpg



http://www.kreativrad.de/img/parts/fahrrad-massanfertigung-koerpergroesse.png

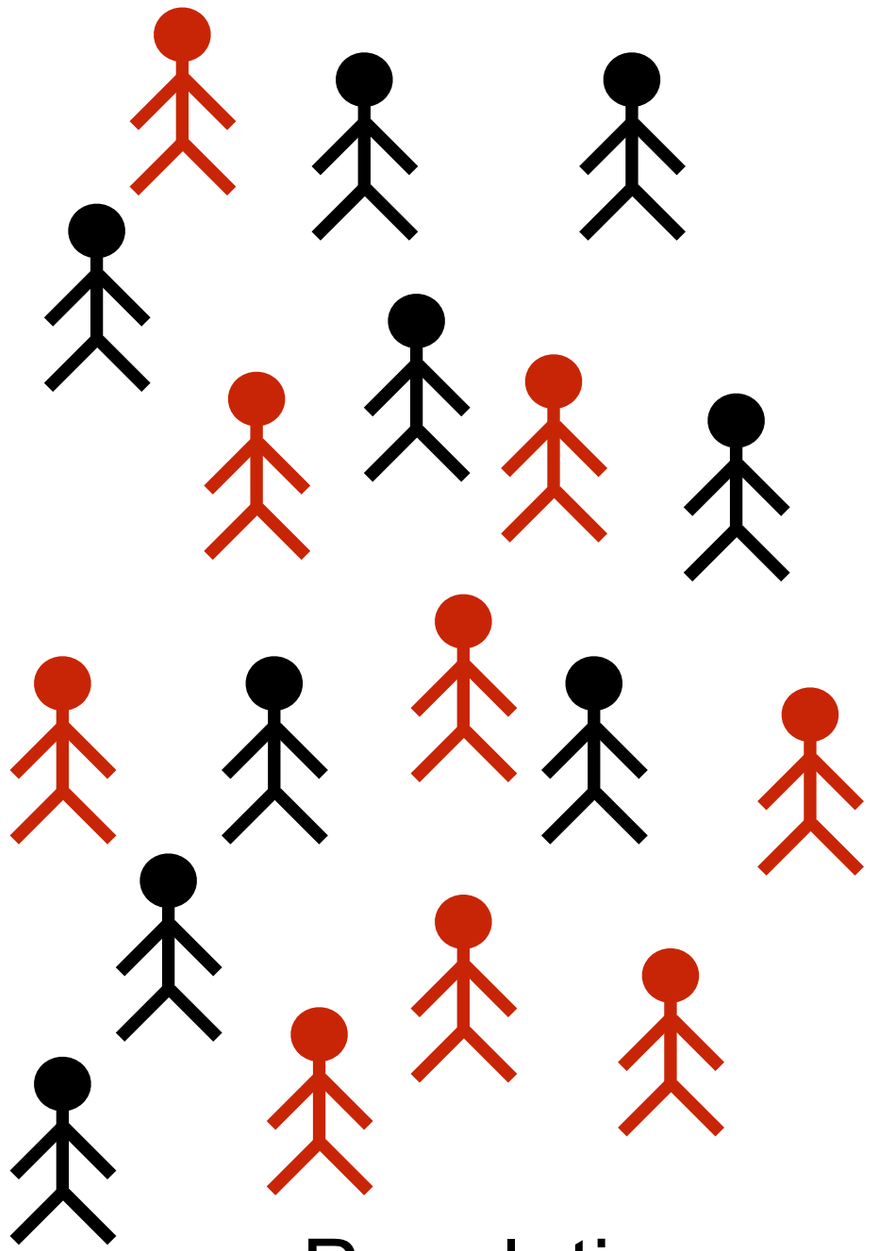
TABELLE

BUNDESLIGA | 2. BUNDESLIGA

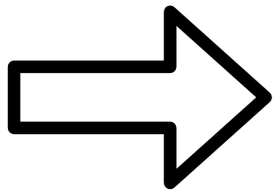
PL	VEREIN	SPIELE	TD	PKT
1	FC Bayern	0	0	0
	Borussia Dortmund	0	0	0
	Schalke 04	0	0	0
	Bayer 04	0	0	0
	VfL Wolfsburg	0	0	0
	Borussia M'gladbach	0	0	0
	Mainz 05	0	0	0
	FC Augsburg	0	0	0
	1899 Hoffenheim	0	0	0
	Hannover 96	0	0	0
	Hertha BSC	0	0	0
	SV Werder	0	0	0
	Eintracht Frankfurt	0	0	0
	SC Freiburg	0	0	0
	VfB Stuttgart	0	0	0
	Hamburger SV	0	0	0
	1. FC Köln	0	0	0
	SC Paderborn	0	0	0

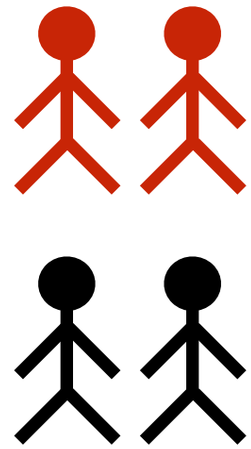
http://www.bundesliga.de/

Subjects



Population

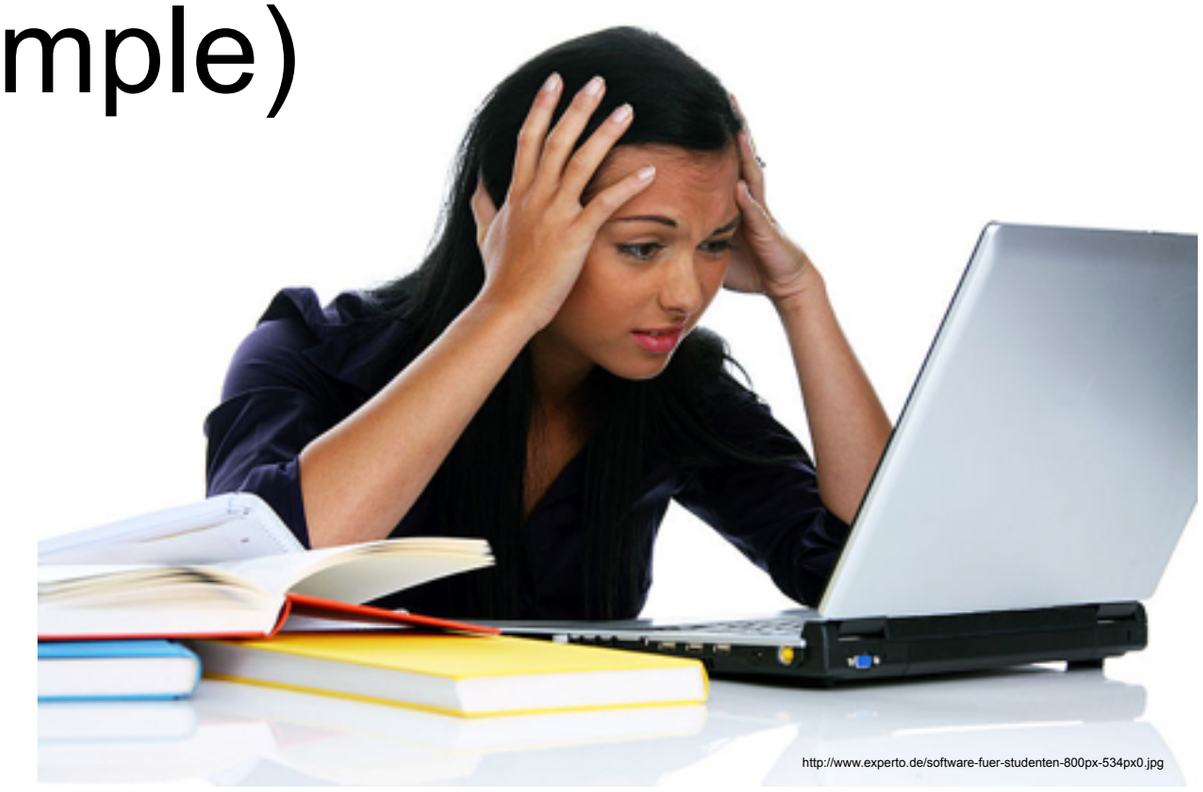




Sample

- Age
- Gender
- Previous knowledge
- Handedness
- Vision
- Education
- Nationality …

Observation Study (Example)



- One independent variable: Participation in tutorials (Yes / No)
  - Assuming participation is voluntary
- One dependent variable: Achieved grade in test
- 108 subjects, 54 “yes”, 54 “no” (to participation question)
- Measurement shows: Grade positively ***correlated*** with tutorial participation
- Beware of ***confounding variables***!

Controlled Experiment

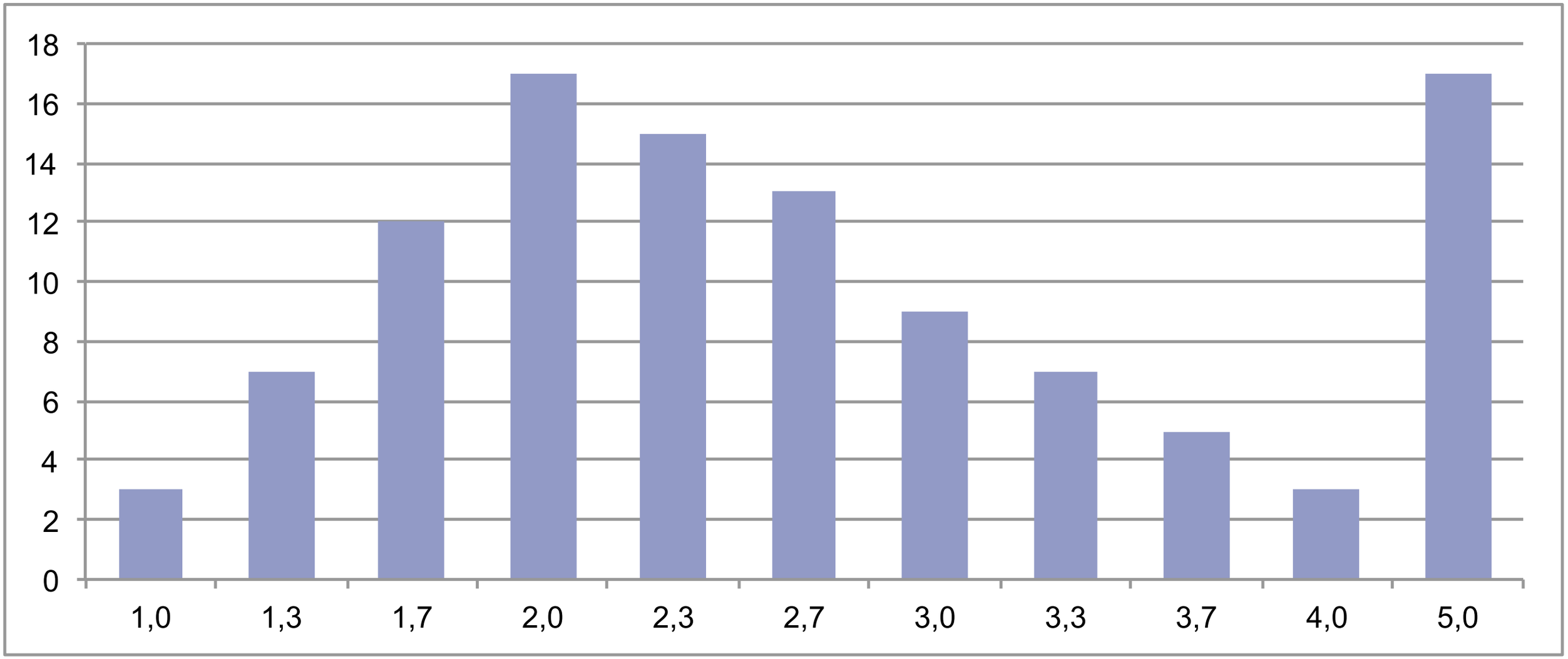
- One independent variable: Participation in tutorials (Yes / No)
  - assigned randomly to subjects !!!
- One dependent variable: Achieved grade in test
- 108 subjects, 54 “participating” condition,
  54 “not-participating” condition
- Measurement: Grade positively ***correlated*** with participation
- Causal relationship established: Participation in tutorials leads to better grade

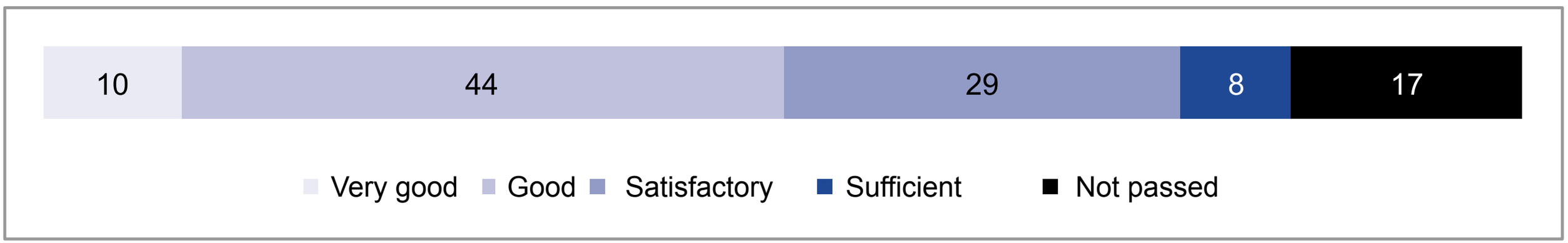
Experiment Design

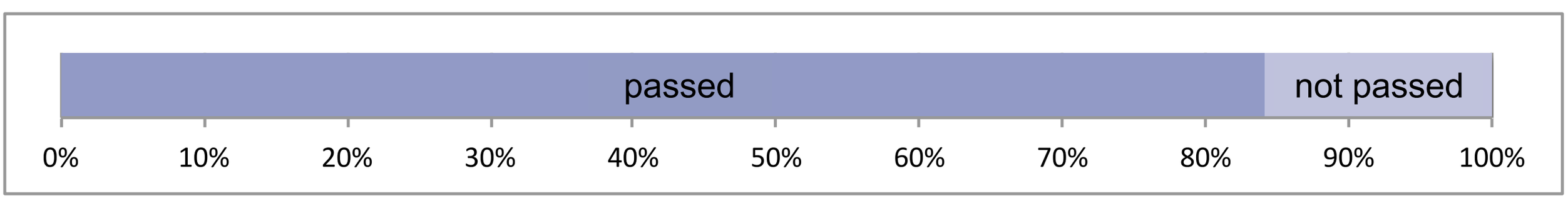
	Int. Design	Analysis	Algebra
Yes	Condition 1	Condition 2	Condition 3
No	Condition 4	Condition 5	Condition 6

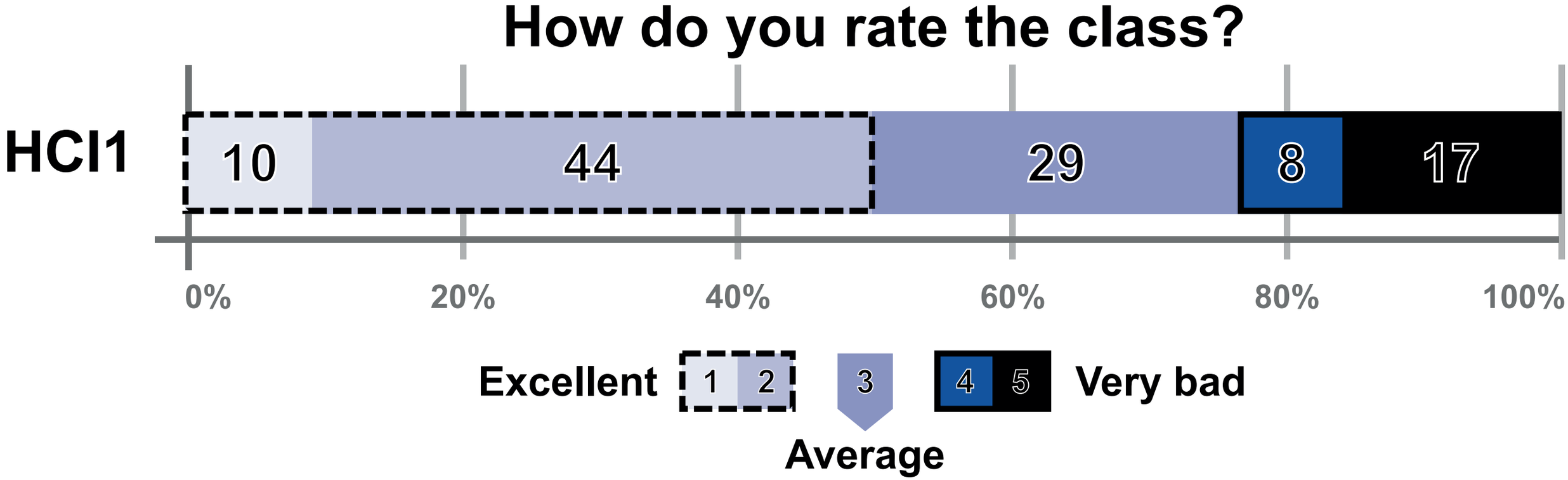
- 2 Variables with 2 resp. 3 values: 2x3 = 6 Conditions
- **within-subjects**: everybody does everything
- **between-groups**: groups, each group does one condition
- Vary the order to avoid **learning** and **fatigue effects**
  - Randomisation
  - Permutation
  - Latin square

Cond. 6	Cond. 1	Cond. 5	Cond. 2	Cond. 4	Cond. 3
Cond. 5	Cond. 6	Cond. 4	Cond. 1	Cond. 3	Cond. 2
Cond. 2	Cond. 3	Cond. 1	Cond. 4	Cond. 6	Cond. 5
Cond. 1	Cond. 2	Cond. 6	Cond. 3	Cond. 5	Cond. 4
Cond. 4	Cond. 5	Cond. 3	Cond. 6	Cond. 2	Cond. 1
Cond. 3	Cond. 4	Cond. 2	Cond. 5	Cond. 1	Cond. 6

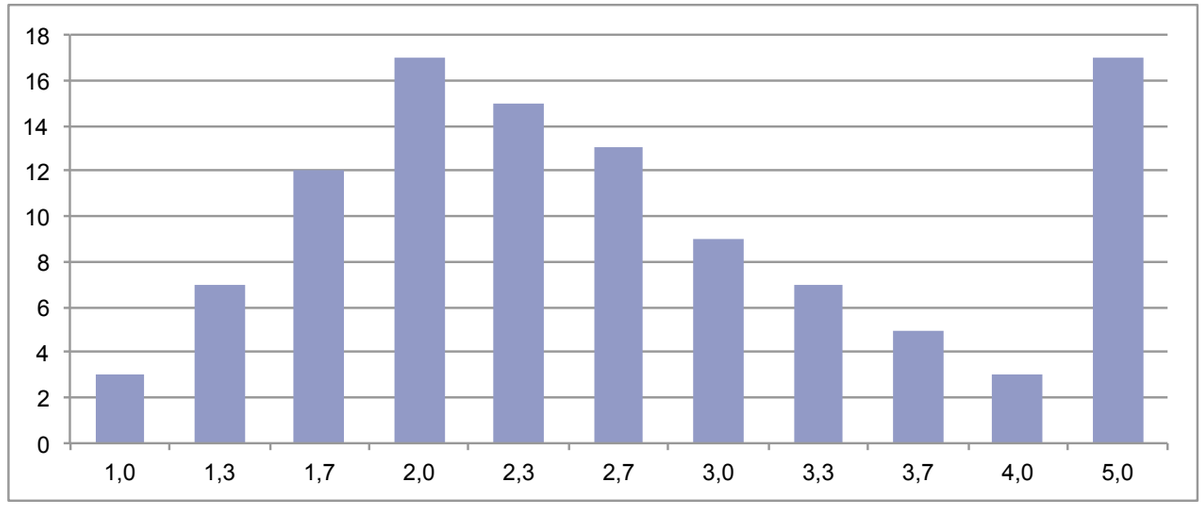

18
16
14
12
10
8
6
4
2
0
1,0
1,3
1,7
2,0
2,3
2,7
3,0
3,3
3,7
4,0
5,0


10
44
29
8
17
Very good
Good
Satisfactory
Sufficient
Not passed


passed
not passed
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


How do you rate the class?
HCI1
10
44
29
8
17
0%
20%
40%
60%
80%
100%
Excellent
1
2
3
4
5
Very bad
Average

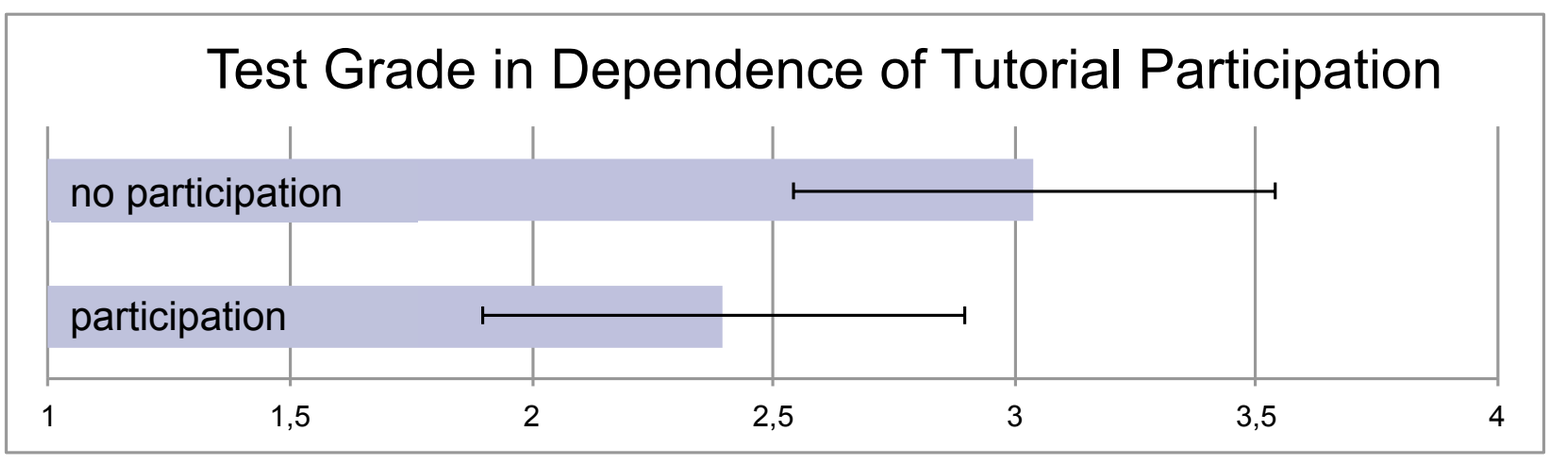
Descriptive Statistics



- nominal data: ***mode*** (2, 4, 5, 5, 5, 5, 5) = 5
- ordinal data: ***median*** (2, 4, 5, 5, 5, 5, 5) = 5
- cardinal data: ***mean*** (2, 4, 5, 5, 5, 5, 5) = 31/7 = 4,42

- standard deviation:
  - median(1,2,3,4,5) = median(3,3,3,3,3) = 3
  - mean(1,2,3,4,5) = mean(3,3,3,3,3) = 3
  - σ(1,2,3,4,5)=1,58
  - σ(3,3,3,3,3)=0,0

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Hypotheses and Significance

- H: Tutorial participants achieve better grades in test.
- H_0: Tutorial participants and non-participants achieve in average the same grades in test.
- Effect size = difference of mean values (unknown in advance)

- Problem: Effect size is not predictable, therefore it is difficult to formulate H in a more precise way

- Trick: Instead of proving H, dis-prove H_0.
  Then H is implicitly proven – independent of effect size.

Significance Tests (Example t-test)

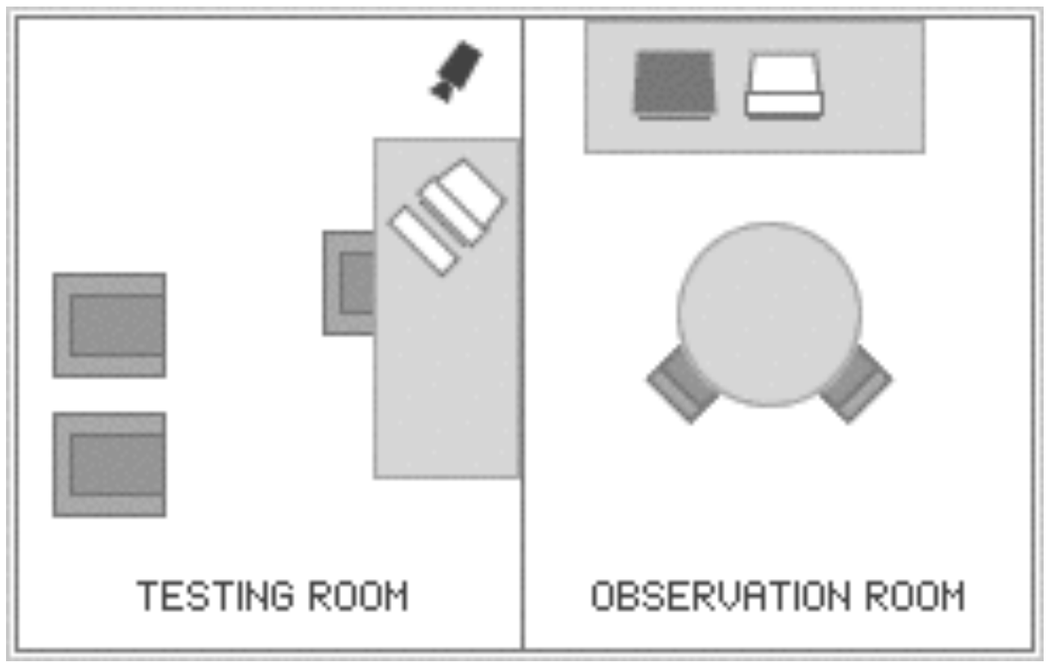
- Input: 2 rows of data
- Output: Probability value *p* between 0 and 1
  - Probability for both rows having in reality the same mean value
- Significance level:
  - Often 0,05 (= 5%)
  - Other values possible: 0,01, 0,001

- If p < 0,05: "***significant difference***" between data rows.

- Different tests für various experiment designs

Field Study vs Lab Study

- External Validitity
- Internal Validitity
- Effort



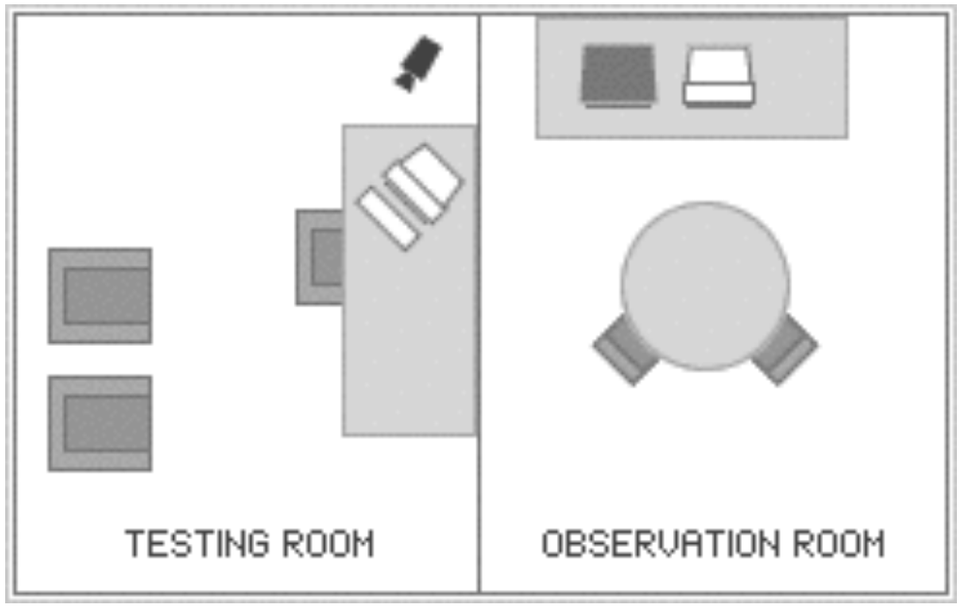
Source: www.xperienceconsulting.com

Field Studies

- Normal activities are studied in normal environment
- Advantages:
  - Can reveal results on user acceptance
  - Allows longitudinal studies, including learning and adaptation
- Problems:
  - In general very expensive
  - Highly reliable product (prototype, mockup) needed
  - How to get observations?
    - Collecting usage data
    - Collecting incident stories
    - On-line feedback
    - Retrospective interviews, questionnaires

Usability Laboratory

- Specifically constructed testing room
  - Instrumented with data collection devices (e.g. microphones, cameras)
- Separate observation room
  - Usually connected to testing room by one-way mirror and audio system
  - Data recording and analysis
- Test users perform prepared scenarios
  - **“Think aloud”** technique
- Problem:
  - Very artificial setting
  - No communication



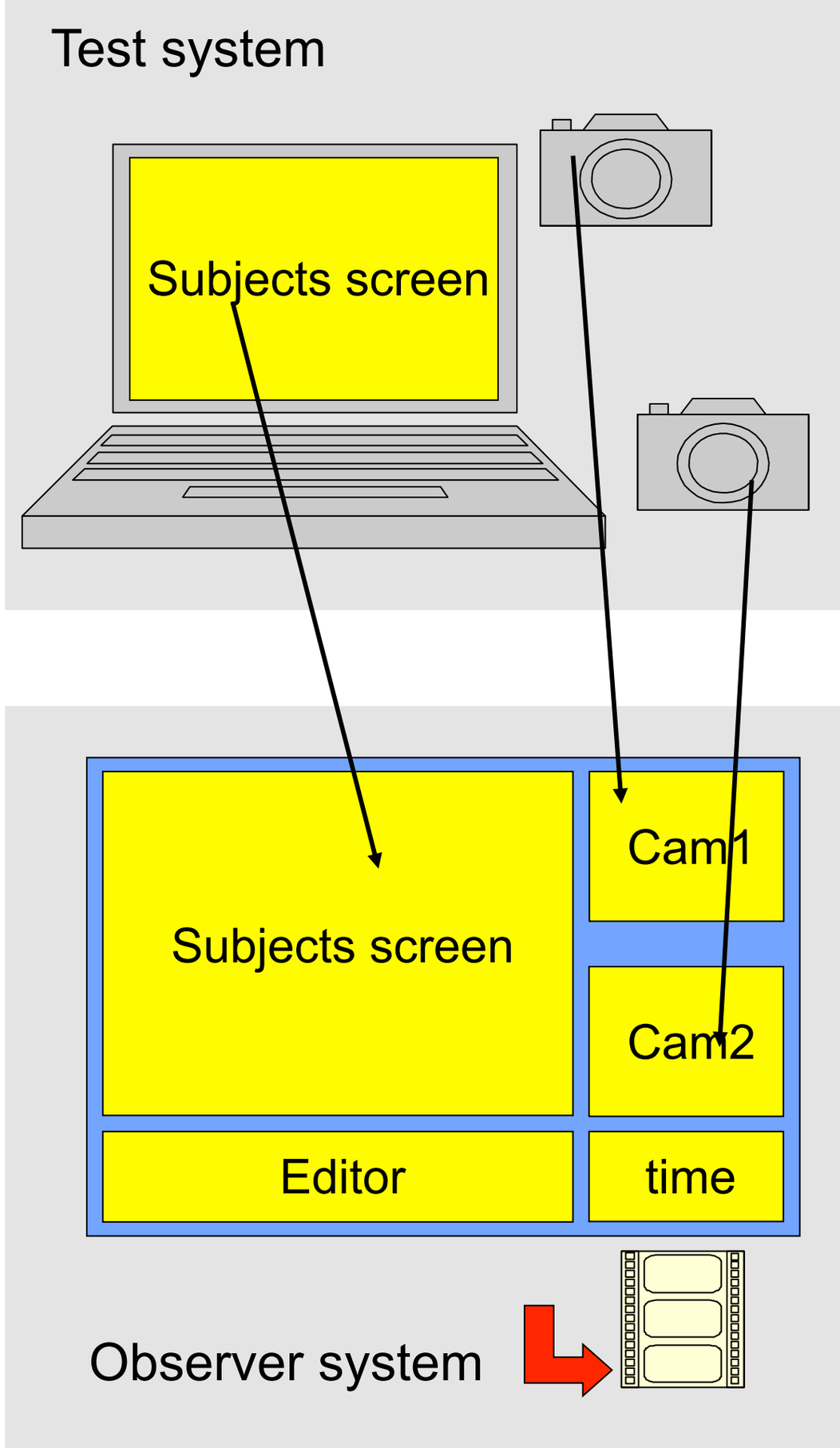
Source: www.xperienceconsulting.com

Poor Man‘s Usability Lab

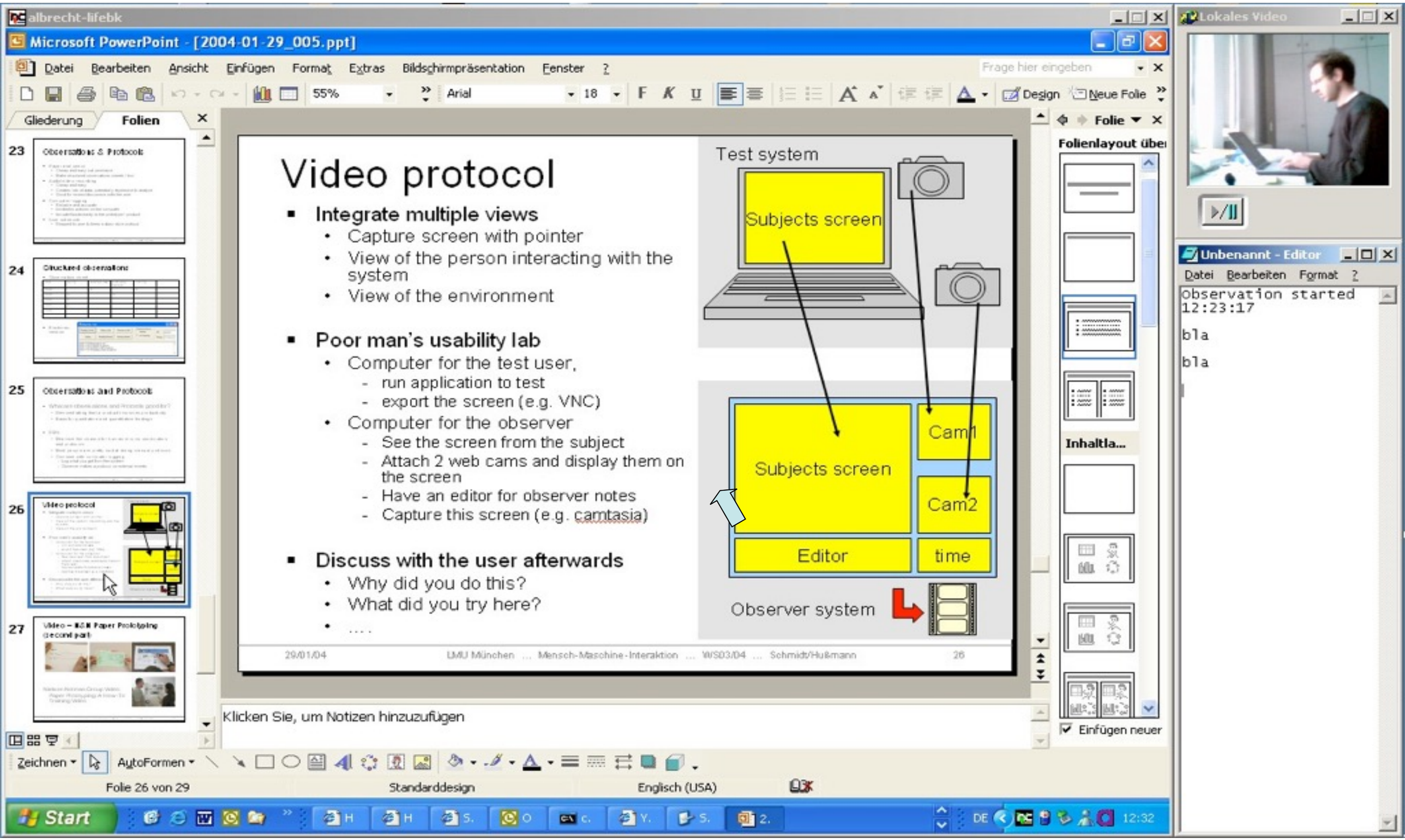
- Goal: Integrate multiple views
  - Capture screen with pointer
  - View of the person interacting with the system
  - View of the environment

- Setup:
  - Computer for the test user,
    - run application to test
    - export the screen (e.g., via VNC)
  - Computer for the observer
    - See the screen of the subject
    - Attach 2 web cams (face and entire user)
    - Display them on the observer‘s screen
    - Have an editor for the observer‘s notes
    - Capture this screen (e.g. QT, Camtasia)

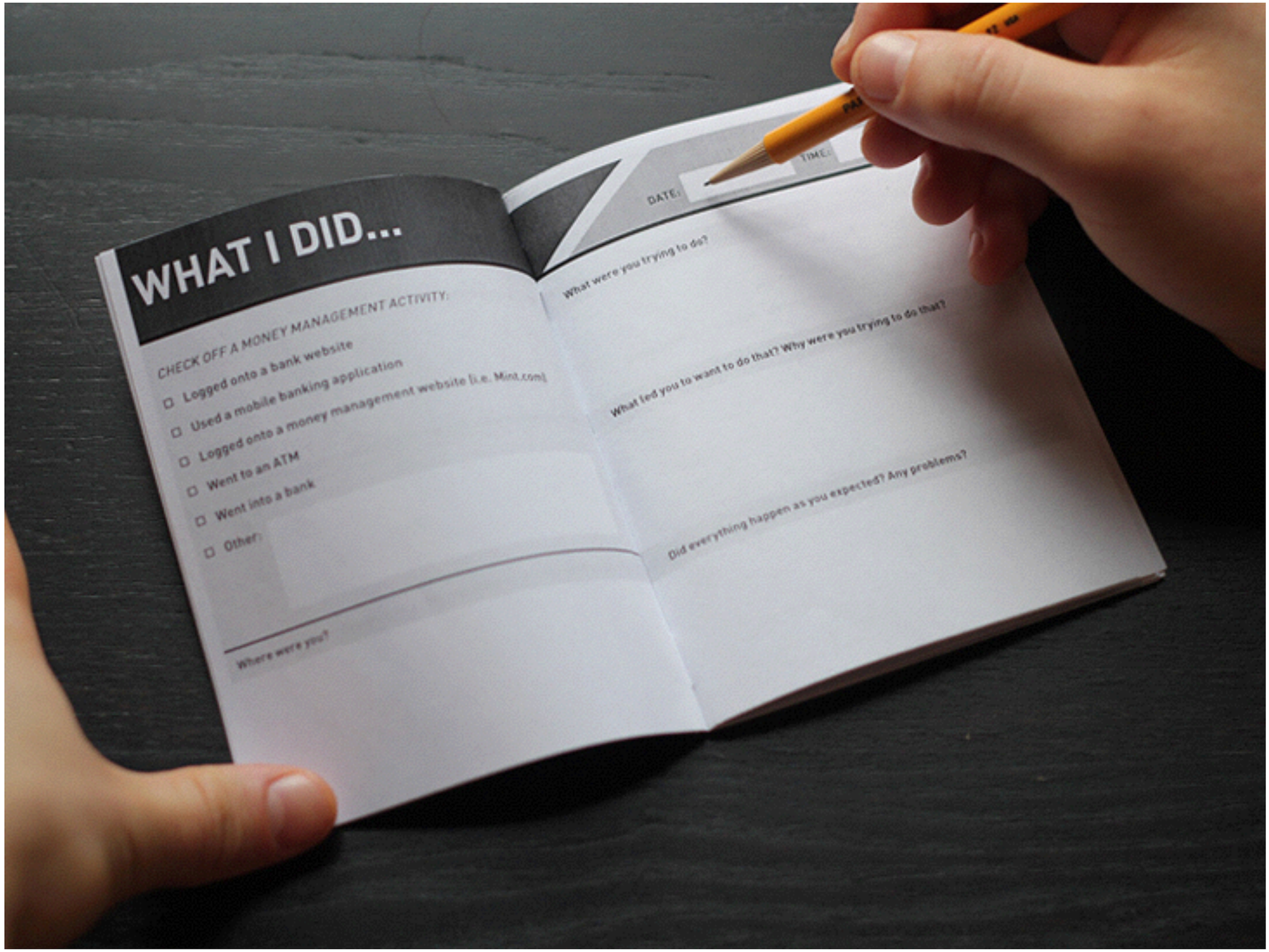
- Discuss with the user afterwards
  - Why did you do this?
  - What did you try here?
  - ….



Screen video



Longitudinal and Diary Studies

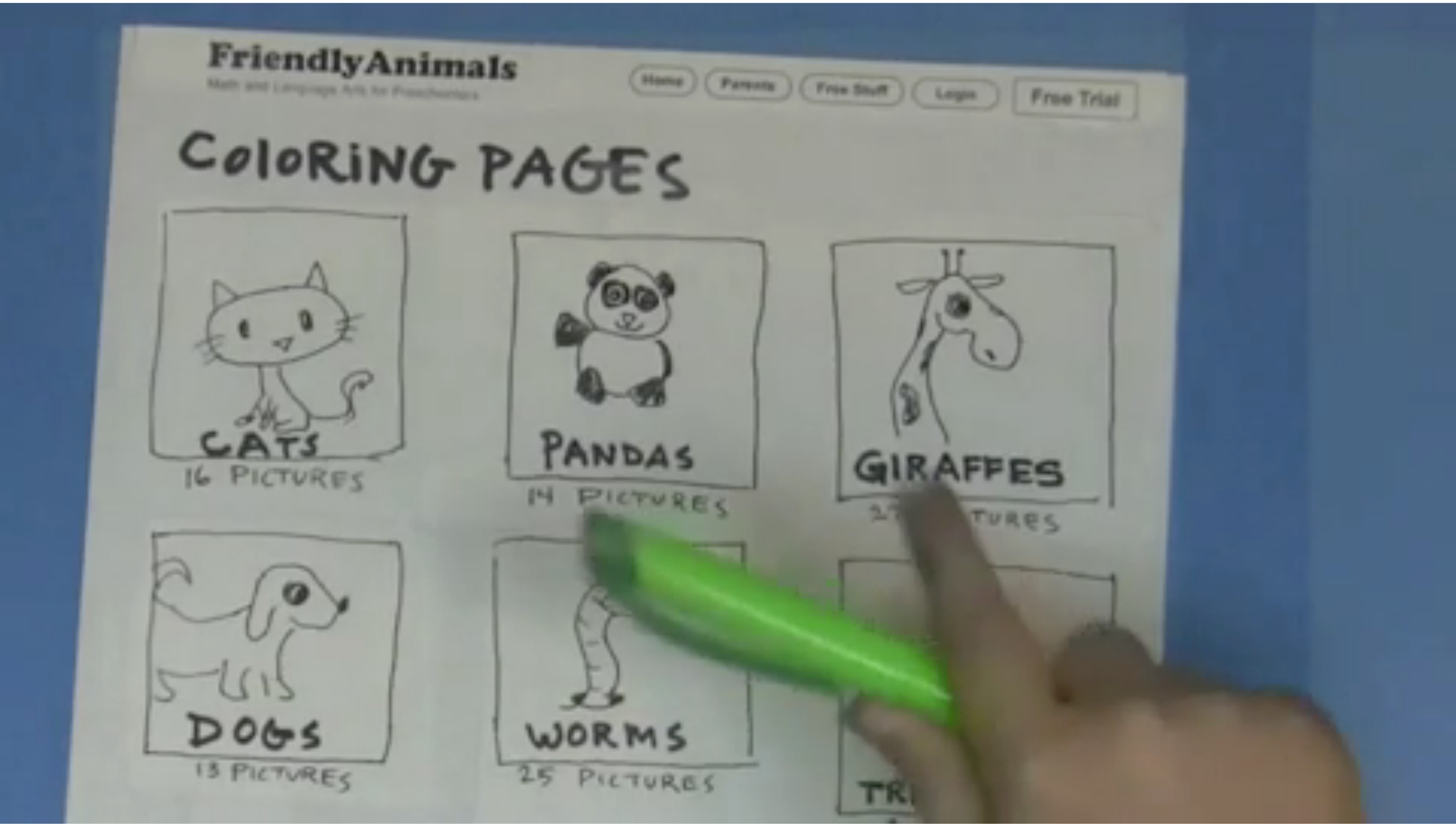


http://www.hcii.cmu.edu/M-HCI/2011/BOA-PlanningTools/images/diary_study.jpg

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough GOMS + KLM	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

Paper Prototype Study



https://www.youtube.com/watch?v=9wQkLthhHKA

Agenda for this class

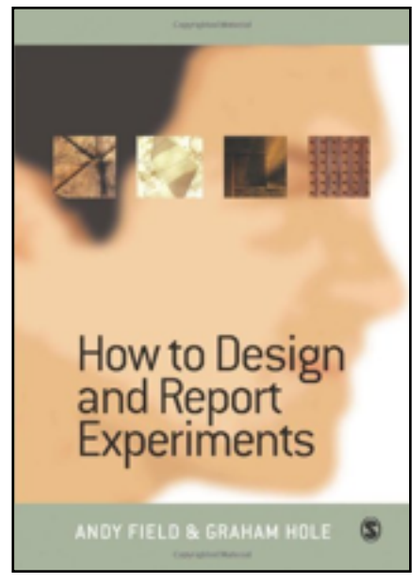
Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough GOMS + KLM	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

References

- Alan Dix, Janet Finlay, Gregory Abowd and Russell Beale: Human Computer Interaction (third edition), Prentice Hall 2003
- Mary Beth Rosson, John M. Carroll: Usability Engineering. Morgan-Kaufman 2002. Chapter 7
- Discount Usability Engineering http://www.useit.com/papers/guerrilla_hci.html
- Heuristic Evaluation http://www.useit.com/papers/heuristic/

- Further Literature
  - Andy Field & Graham Hole: How to design and report experiments, Sage
  - Jürgen Bortz: Statistik für Sozialwissenschaftler, Springer
  - Christel Weiß: Basiswissen Medizinische Statistik, Springer
  - Lothar Sachs, Jürgen Hedderich: Angewandte Statistik, Springer
  - various books by Edward R. Tufte

– video on next slide by Eric Shaffer, Human Factors Inc. http://www.youtube.com/watch?v=bminUlAu47Q



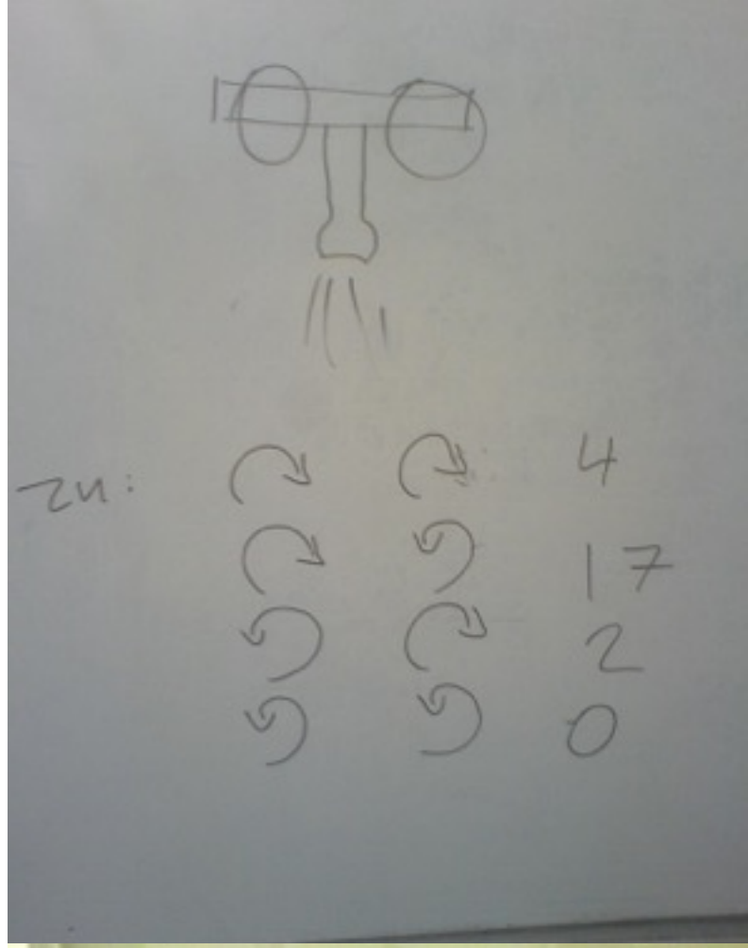
http://www.amazon.de/dp/0857028294


HOT


COLD

Intuitive Interfaces?

- Given: old style water faucet
  - 2 valves, 1 outlet
  - Cylindrical, next to each other
  - Left warm, right cold



- Question: In which direction does each valve close?

- Homework: find such faucets, determine which are „intuitive“ and why (not)