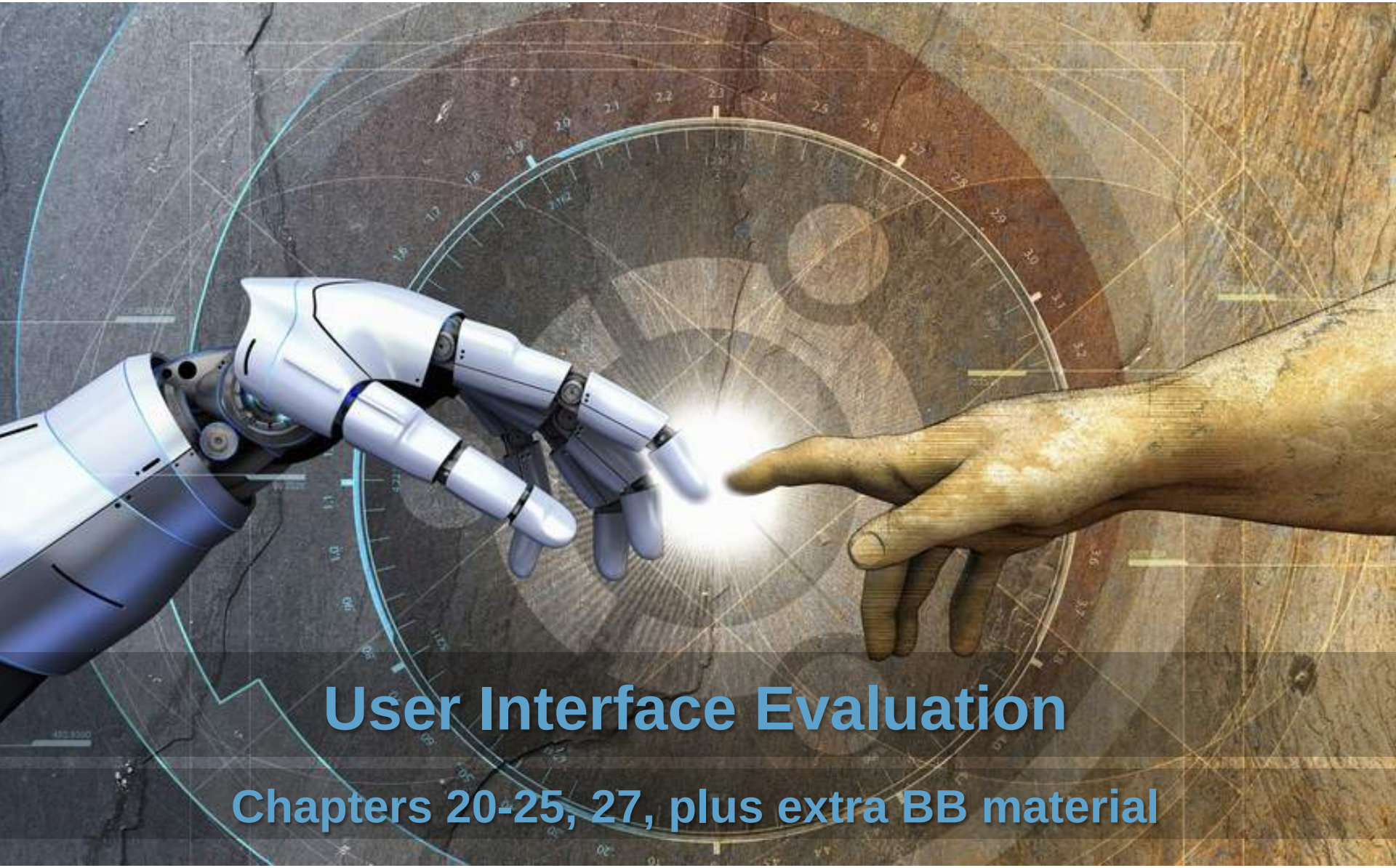
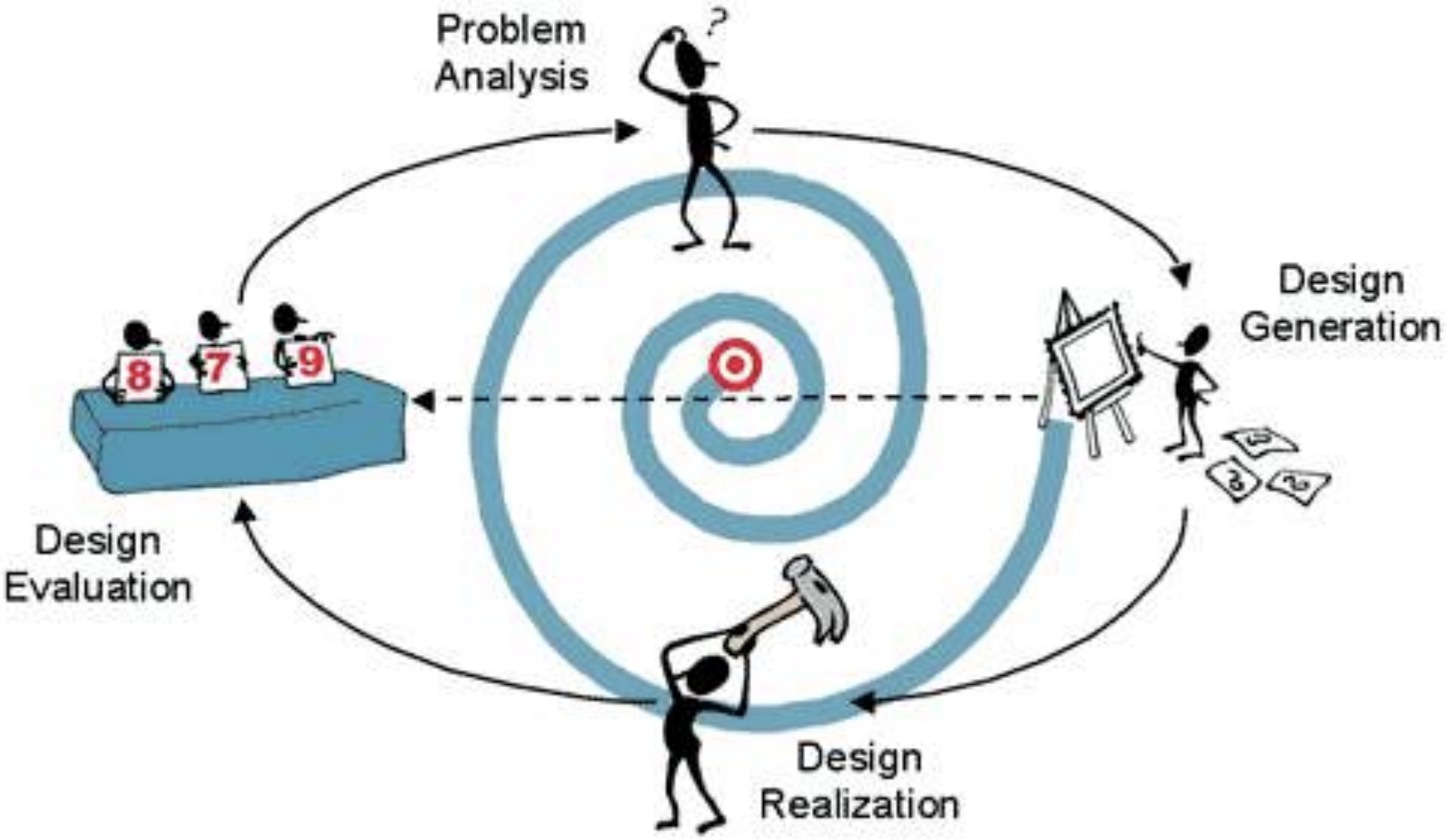


User Interface Evaluation

Chapters 20-25, 27, plus extra BB material

Lora Aroyo
Web & Media Group


Problem Analysis
Design Generation
Design Realization
Design Evaluation
8
7
9

Evaluation methods

- Expert analysis, such as
  - Cognitive walkthrough
  - Heuristic inspection
- User study
  - Involving users in the evaluation process

- There is a role for both types of methods

Overview

- Example user studies
- Purpose of a user study
- User study design
  - Terminology
  - Types of experiments
  - Reliability, validity and generalization
- Pragmatic aspects of setting up a user study



USER STUDY EXAMPLES

iPhone Study

Research firm User Centric has released a study that tries to gauge how effective the iPhone's unusual on-screen keyboard is. The goal is certainly a noble one, but I can't say that the survey's approach results in data that makes much sense.

User Centric brought in twenty owners of other phones--half who had ones with QWERTY keyboards, and half who had ordinary numeric phone keypads. None were familiar with the iPhone. The research involved having the test subjects enter six sample text messages with the phones they already had, and six with an iPhone.

- Setting:
  - iPhone's touch keyboard compared to *conventional QWERTY* & *numeric phone keyboards*
- Hypothesis:
  - texting potentially problematic for new iPhone users
- Goal:
  - how easy it is for conventional mobile phone users to text with iPhone
- Task:
  - text using both conventional phones and iPhones

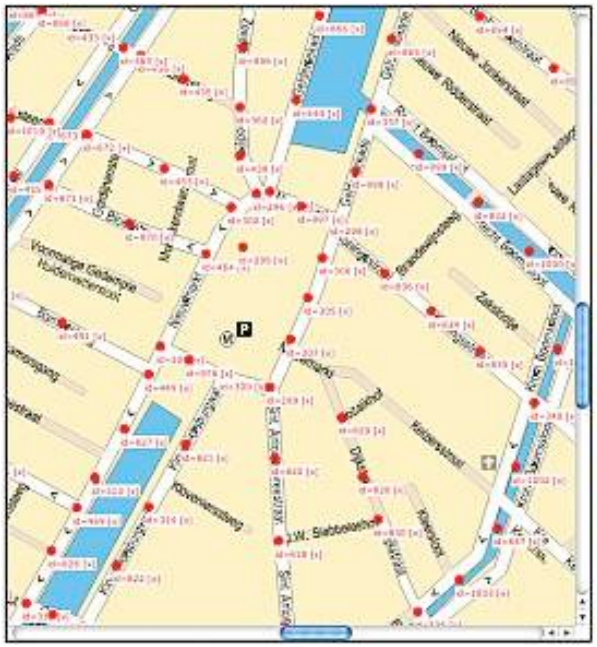
iPhone Study: Result

- texting on iPhone took twice as long as texting on numeric or QWERTY phones
- to be expected since users had much more experience with their own phones
- interesting result:
  - iPhone did same or better for multi-tap users

Study was actually about …

- initial adoption of iPhone keyboard compared to users' current phones
- not survey → study (interviews users typing, timed)
- multitap (non-QWERTY) users did same/better with iPhone → maybe easier adopting iPhones
- expert iPhone texters → to switch to a QWERTY phone → see if similar difference in typing efficiency

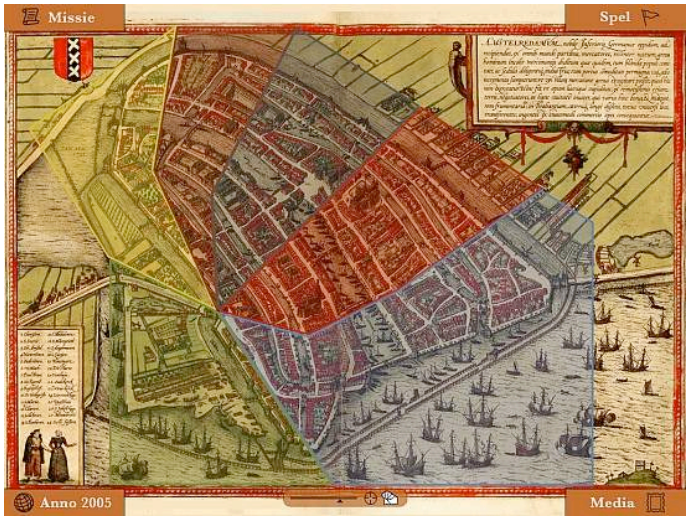
Mobile-pilot in medieval Amsterdam



Mobile in medieval Amsterdam

- ‘Mobile history lesson’: mobile learning game
- mobile phones and GPS-technology (UMTS)
- the mobile game experience fits traditional curriculum

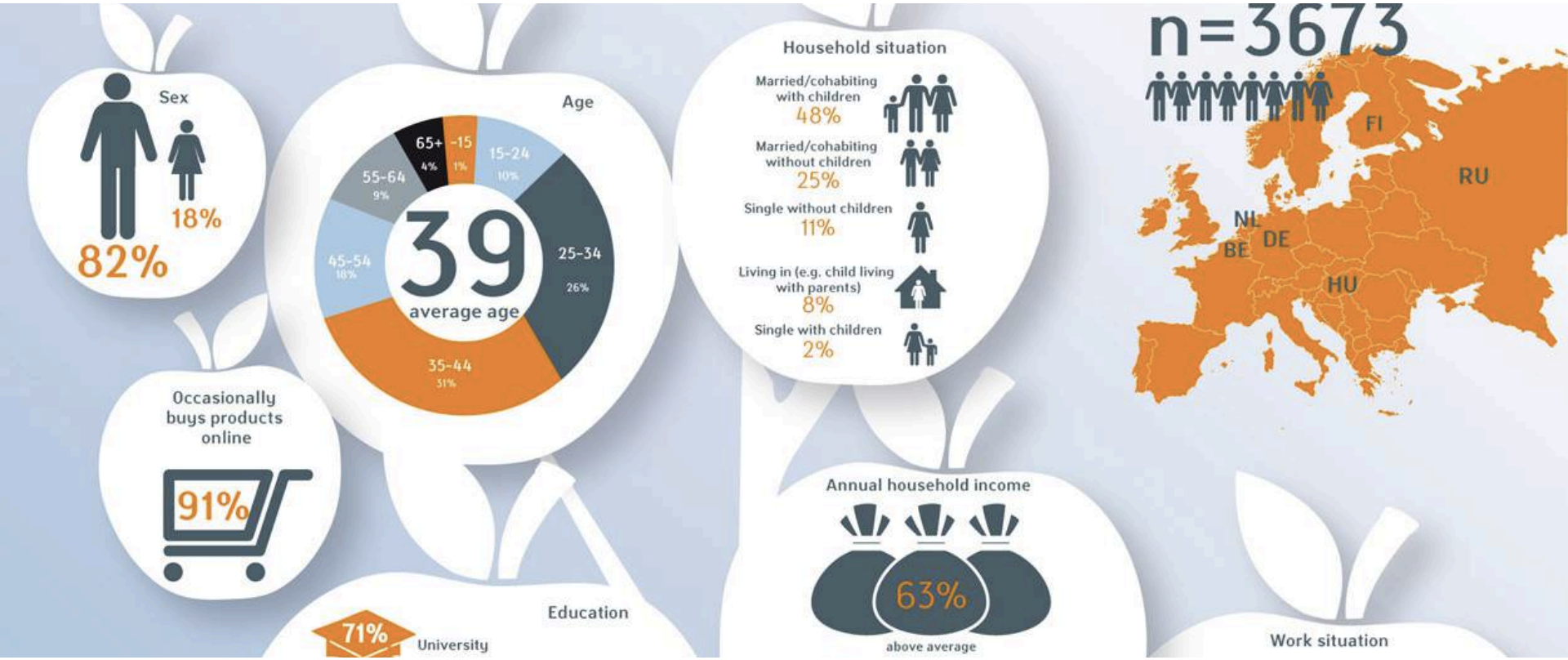
- 10 school classes from 5 different schools
- 467 students in total
- age 12-14
- June 2007



http://freq1550.waag.org/index.html
http://freq1550.waag.org/clips/journaalclip.html

Mobile-pilot in medieval Amsterdam

- Goal:
  - examining technology-supported location-based experience
  - actively experiencing history with immersing (location-based) game & creation of own pictures, audio/video adds to understanding of history
- Hypotheses:
  - Enhances communication & collaboration skills?!
  - Enhance educational abilities, e.g. interpreting historical sources
- surveys for students & supervisors & test scores
  - compared to scores of students who hadn't played the game, but just learnt from a history book



PURPOSE OF A USER STUDY

Usability Testing

- a means for measuring how well people can use some human-made object
  - e.g., a web page, a computer interface, a document, or a device
- for its intended purpose
  - i.e., usability testing measures the usability of the object

We need testing because

- Can't tell how good UI is until?
  - people use it!
- Testing against different app/tool/ interface that promises the same functionality
- Hard to predict what real users will do

Usability Goals

- Effective & efficient & safe to use
- Have good utility
- Easy to learn & to remember how to use

- How long should it take & how long does it actually take:
  - to use a VCR to play a video?
  - to use a VCR to pre-record two programs?
  - to use an authoring tool to create a website?

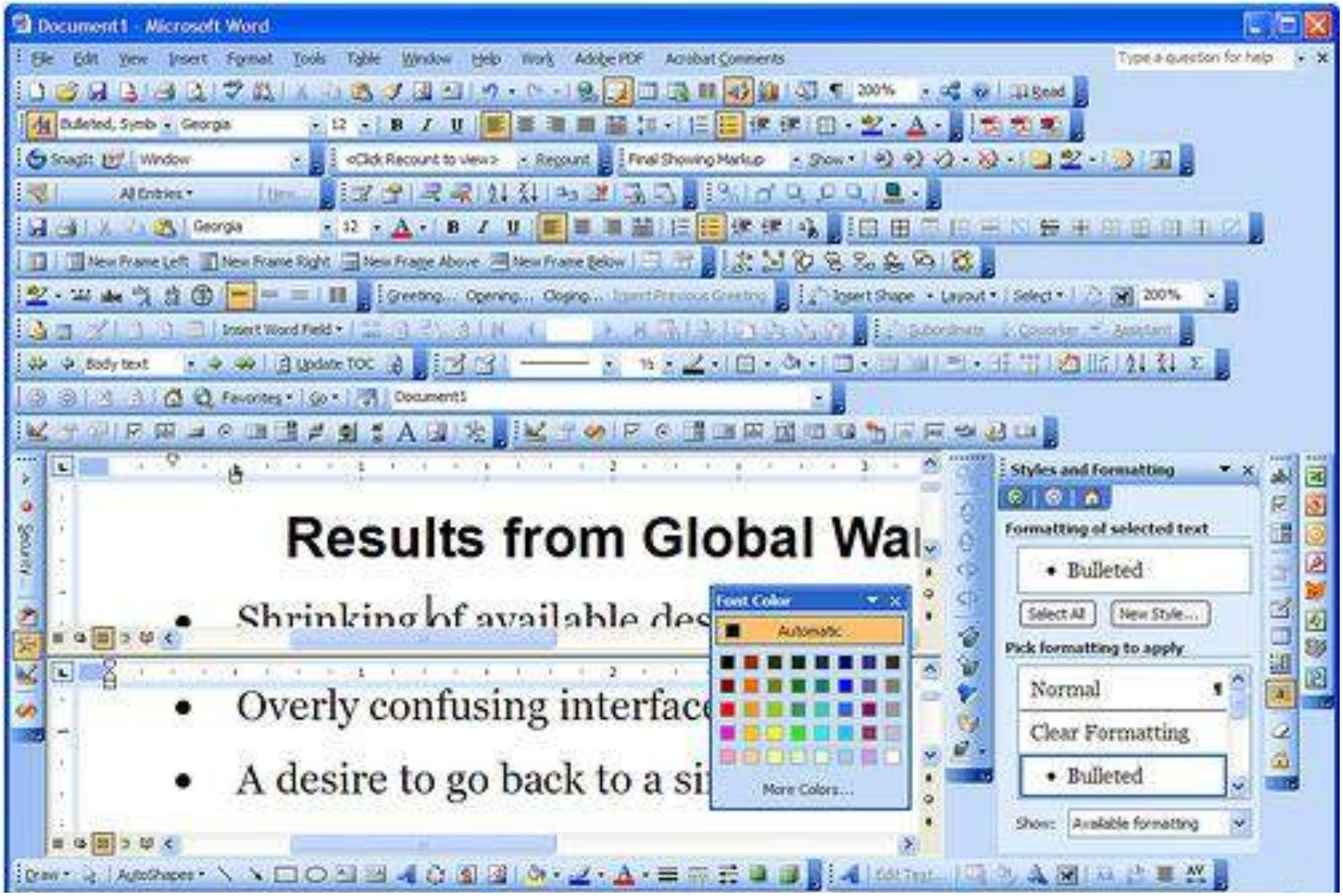
User Experience Goals vs. Usability Goals

- Satisfying, e.g., support creativity, emotionally fulfilling, rewarding
- Fun & Enjoyable & Entertaining
- Helpful & Motivating
- Aesthetically pleasing

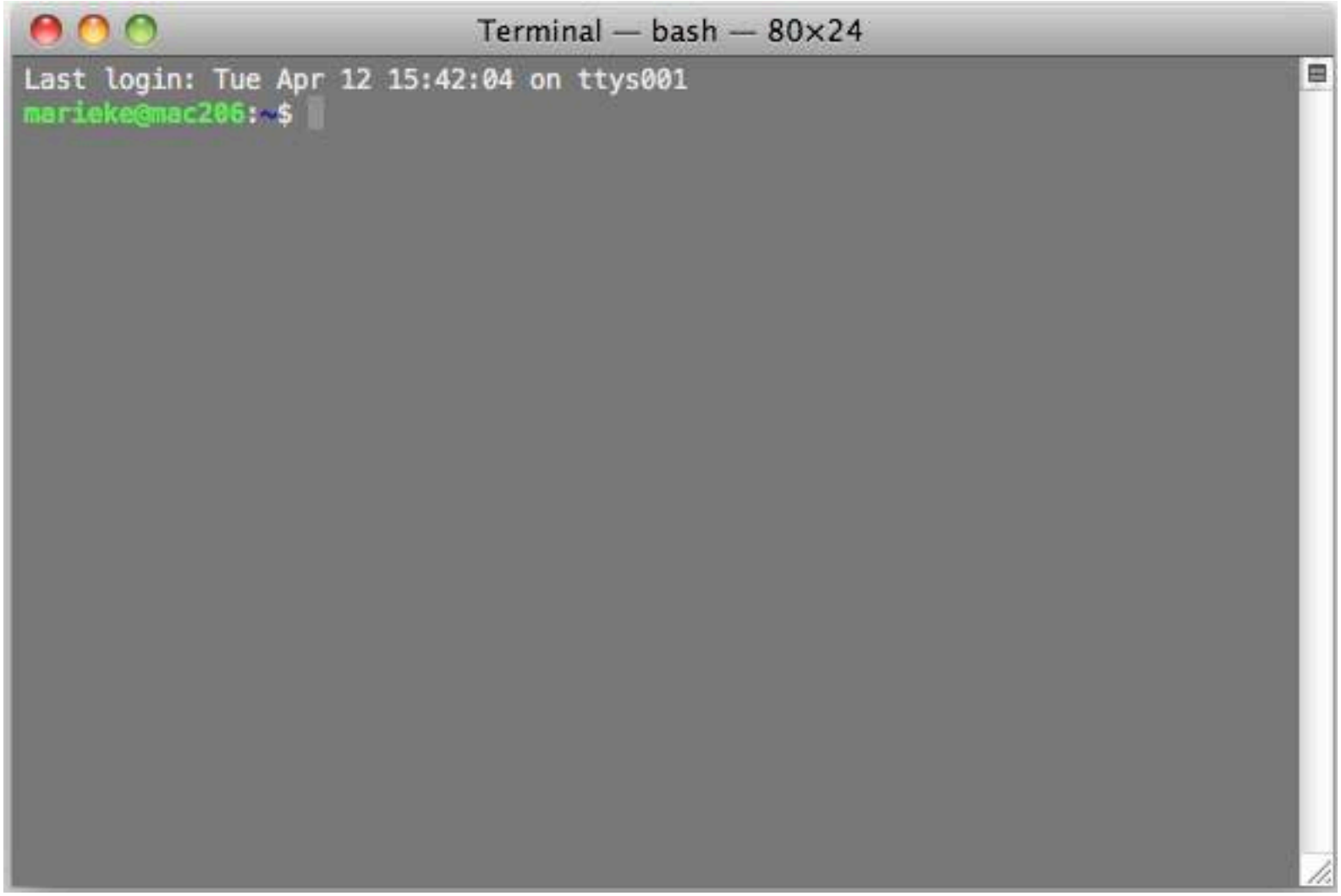
Characteristics to be Tested

- for the different functionalities identify:
  - usage
    - main tasks, scenarios, constraints
  - interaction methods or paradigms
    - communication, selection, commands, …
    - 'panic' fallback option
    - page layout, e.g. size, alignment, fonts
    - clickable elements
  - technologies
    - required for most important activities
    - platform constraints

Too many options



Too few?


Terminal — bash — 80×24
Last login: Tue Apr 12 15:42:04 on ttys001
marieke@mac206:~$

Potential Problems Identification

- user population aspects
  - limitations for children, elderly, casual vs. expert users etc.
- implementation issues
  - restrictions of the current prototype
  - menu structure often shaped by features developer wanted in there, not user needs

Example: TV Guide on your Mobile

- physical aspects, e.g.
  - screen size, buttons available to perform a main task
    - to log in the settop box
    - to view a TV program with corresponding controls
    - to view a TV guide overview

Example: TV Guide on your Mobile

- internal consistency, i.e.,
  - within mobile software
    - main menu and explanations (context-sensitive help) always available
    - colors and layout consistency
    - navigation controls consistency
    - input (query, browse) and output (results presentation) consistency

Example: TV Guide on your Mobile

- external consistency, i.e.,
  - with desktop software
    - Online TV Guide interface consistent with the Mobile interface
    - Social Application interface consistent with the TV Guide interface
    - Social Mobile interface consistent with general mobile interfaces

What should I study?

- Time with a Task
  - How long does it take people to complete basic tasks?
    e. g., find something to buy, create a new account, order an item
- Accuracy
  - How many mistakes did people make?
  - Were they fatal or recoverable with the right information?
- Recall
  - How much does the person remember afterwards or after periods of non-use?
- Emotional Response
  - How does the person feel about the tasks completed?
  - Confident? Stressed?
  - Would the user recommend this system to a friend?

Goals of HCI evaluation

- Depends on the stage of a project:
  - Ideas and concepts
  - Designs
  - Prototypes
  - Implementations
  - Products in use
- Differenciate between assessing learnability or interaction

  ⇨ train the user before the tasks?
- Approaches
  - Formative evaluation
    - » Throughout the design
    - » Helps to shape a product
  - Summative evaluation
    - » Quality assurance of the finished product

Formative vs. summative evaluation

Qualitative:
- Get "non-measurable" feedback
- General insight

- Used to
  - Find problem areas
  - Find conceptual errors
  - Find missing functionality

Quantitative:
- Measure performance
- Generate statistical data

- Used to
  - Verify performance benefits of new input/output devices or interaction techniques
  - Determine differences between user groups

most useful for formative evaluation

most useful for summative evaluation

Formative Evaluation:
- Throughout the design
- Helps to shape a product

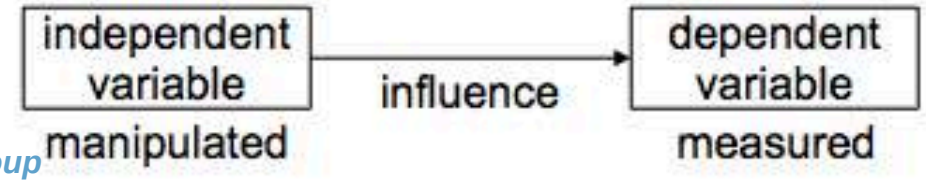
Summative Evaluation:
- Quality assurance of the finished product



USER STUDY DESIGN

Terminology: Independent vs. dependent variables

- **Independent** variables
  - Manipulated by the experimenter
  - Conditions under which the tasks are performed
  - The number of different values used is called ***level***, e.g.
    - » Font can be *Arial* or *Times* (2 levels)
    - » Background can be blue, green, or white (3 levels)
- **Dependent** variables
  - Affected by the independent variables
  - Measured in the user study
  - Objective values: e.g. time to complete a task, number of errors, etc.
  - Subjective values: ease of use, preferred option
  - They should only depend on the independent variables (conditions)



Terminology - extended

- **Response variables** = dependent variables
  - usually cannot be directly controlled
  - *What do I observe? - outcomes of experiment*
- **Factors** = independent variables
  - *What do I change? - variables we manipulate in each condition*
  - to determine its relationship to an observed phenomenon, i.e. dependent variable
- **Controlled variables** = kept constant to prevent their influence on the effect of the independent variable on the dependent
  - *What do I keep the same?*
- **Levels** = possible values for independent variables

How should I study it?

- Choose a dependent variable
- Manipulate the independent variable
- Try to avoid interference from other variables
- Measure results

Example

Measure the influence of different fertilizer quantities on plant growth

- IV= the changing factor of the experiment
  - IV_1 = amount of fertilizer used
- DV = the factors that are influenced in the experiment
  - DV_1 = growth in height
  - DV_2 = mass of the plant
- CV = the factors that would influence the dependent variable if not controlled
  - CV_1 = type of plant
  - CV_2 = type of fertilizer
  - CV_3 = amount of sunlight the plant gets
  - CV_4 = size of the pots

User study outcomes (Example)

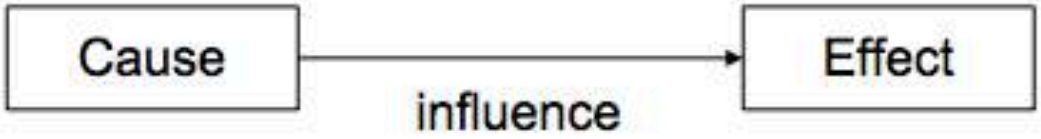
- Users are quicker using version A than using version B
- Users make 10% less errors when using version X than when using version Y
- 90% of the users can complete the transaction using version Y in less than 3 minutes
- On average users will be able to buy a ticket using version A in less than 30 seconds

Empirical scientific questions

- Base rates: How often does Y *occur*?
  - Requires measuring Y
- Correlations: Do X and Y *co-vary*?
  - Requires measuring X and Y
- Causes: Does X *cause* Y?
  - Requires measuring X and Y, and manipulating X
  - Also requires somehow accounting for the effects of other independent variables (confounds)!

Cause and effect

- Why do scientists measure things?
  ⇨ Find causal links between variables, e.g. smoking ⇨ cancer



- Criteria that need to be met to infer cause and effect (Mill):
  1. Cause has to precede effect
  2. Cause and effect should correlate
  3. All other explanations of the cause-effect relationship must be ruled out

- Only way to infer causility:
  - Two controlled situations
    1. Cause is present (*experimental condition*)
    2. Cause is absend (*control condition*)
  - Otherwise the situations have to be identical!

What do I expect to find? (hypotheses)

- Consist either of:
  - a suggested explanation for a phenomenon
  - a reasoned proposal suggesting a possible correlation between multiple phenomena
- one can test a scientific hypothesis
- generally hypotheses are based on previous observations or on extensions of scientific theories
- Testable using observation or experiment
  - all evidence must be **empirical**, or **empirically** based
    - evidence or consequences that are observable by the senses

Hypothesis

- Prediction of the result
- States how a change in the independent variables will effect the measured dependent variables
- By doing an experiment, the hypothesis is either proved or disproved
- Null hypothesis predicts that independent variables do not have any effect on the dependent variables
- Formulate hypotheses BEFORE running the study!

Example Hypotheses

- A *mobile museum guide* helps users in *finding their way* in a museum
- A *search field generally available* in the online Museum Tour Wizard, helps users *find quicker the artworks* to include in their tours
- *Mobile-based identification* for digital TV allows users to *login unobtrusively* to their home STB (set-top box)
- The *wiki review/contribution policies* in the tagging interface *increases the trust and motivation of users*.

Can my experimental design be meaningfully analyzed?

- Reliability
- Validity
- Generalizability

Reliability

Good experiments should yield:

- A ‘true score’ for that which we were aiming to measure
- A ‘score for other things’ we are measuring inadvertently
- Systematic (non-random) bias
- Random (non-systematic) error

Maximizing reliability

- Precise, unambiguous and objective definition of what is being measured.
- Not always easy!
  - Easy examples:
    - » Memory ⇨ # items recalled
  - Hard example: measuring effect of frustration on children's agression
- Solutions
  - Definition by consensus
    - » Find canditates for agressive activities (e.g. through observations)
    - » Independent judges rate agression of activities
  - Operational definition
    - » Experimentor defines agressive behavior as X, Y, Z for the purpose of this study
    - » Whether one agrees to the definition or not, at least the results are true for X, Y, Z

Validity

- Concerns the relationship between concept and indicator
  - Measuremens show what they are intended to show
- Internal validity
  - Measurements are accurate
  - Measurements are due to manipulations, not caused by other factors
  - Precondition:
    - » Good experimental design
- External validity
  - Findings are representative of humanity
  - Not only valid in experiment setting
  - Precondition:
    - » Good judgement and sometimes intuition

Internal Validity

- manipulation of independent variable is a cause of change in dependent variable (measurements are accurate, as well as the experimental design)
  - requires removing effects of confounding factors (controlled variables)
  - requires choosing a large enough sample size, so the result couldn't have happened by chance alone

External Validity

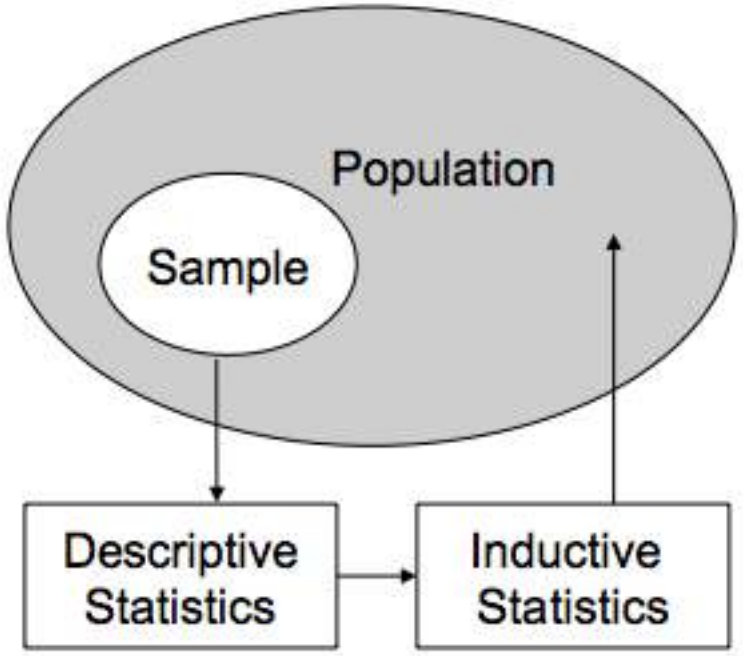
- Experiment possess external validity
  - if the experiment's results hold across different experimental settings, procedures and participants (experiment be replicable)
  - if results of the experiment generalize to the larger population (real world situations)
  - no study "has" external validity by itself!

- The most common loss of external validity:
  - experiments using human participants often employ small samples obtained from a single geographic location
  - one can not be sure that any results obtained would apply to people in other geographic locations

Generalizability

Example threat: overusing the same participant group

- Psychology findings are biased through extensive use of 1st-year students as test persons



Control vs. Randomization

- ***Control**:* holding a variable constant for all cases
  - Lower generalizability of results
  - Higher precision of results
- ***Randomization**:* allowing a variable to randomly vary for all cases
  - Higher generalizability of results
  - Lower precision of results
- ***Randomization within blocks:*** allowing a variable to randomly vary with some constraints
  - Compromise approach

How to gather results?

- Observational
  - if done unobtrusively, yields reliable results
  - time-consuming & cause-effect is not always clear
- Quasi-experimental
  - test group and control group are not randomly allocated, but on basis of pre-existing differences
  - pre-existing differences may influence results
- Experimental
  - participants are randomly assigned to groups

Experimental vs. observational methods

Two approaches to answering research questions (RQ)

1. **Observational** (= correlational) methods:
   Observe what naturally happens in the environment without interfering
2. **Experimental** methods:
   Manipulate some aspects and observe the effects

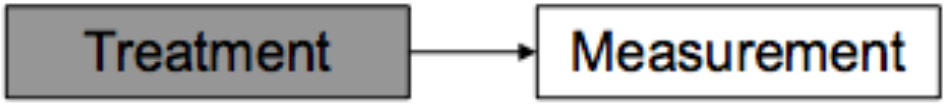
	Experimental	Observational
Pros	• Isolate and control variables ⇨ allow causal statements	• Natural setting: observe how people behave normally
Cons	• Danger of artificial situations ⇨ people might behave differently	• Variables are not isolated • Time consuming

Quasi experiment - motorcycle example

- RQ: Does daytime headlight use make motorcyclists more detectable?
- Dependent variable: number of accidents
- Experimental design:
  - Randomly allocate large group of motorcyclists to two groups
    - » Experimental group uses headlight during daytime
    - » Conrol group does not use headlight during daytime
  - Ethical reasons against this allocation!
- Solution: Quasi-experimental design:
  - Find motorcyclists with different preferences
  - Pre-existing difference (⇨ group threat):
    Other factors related to the preference for/against headlights can influence the dependent variable, e.g.
    - » Older machines
    - » Different safe-conscious levels

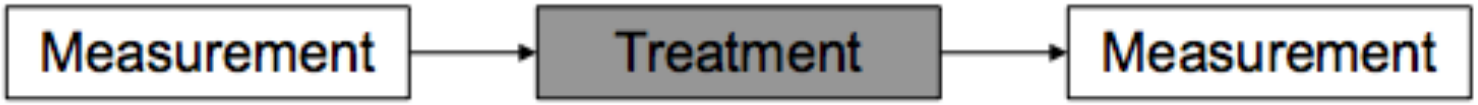
Types of quasi experimental design

1. One group post-test design



- No baseline against which to compare

2. One group pre-test/post-test design



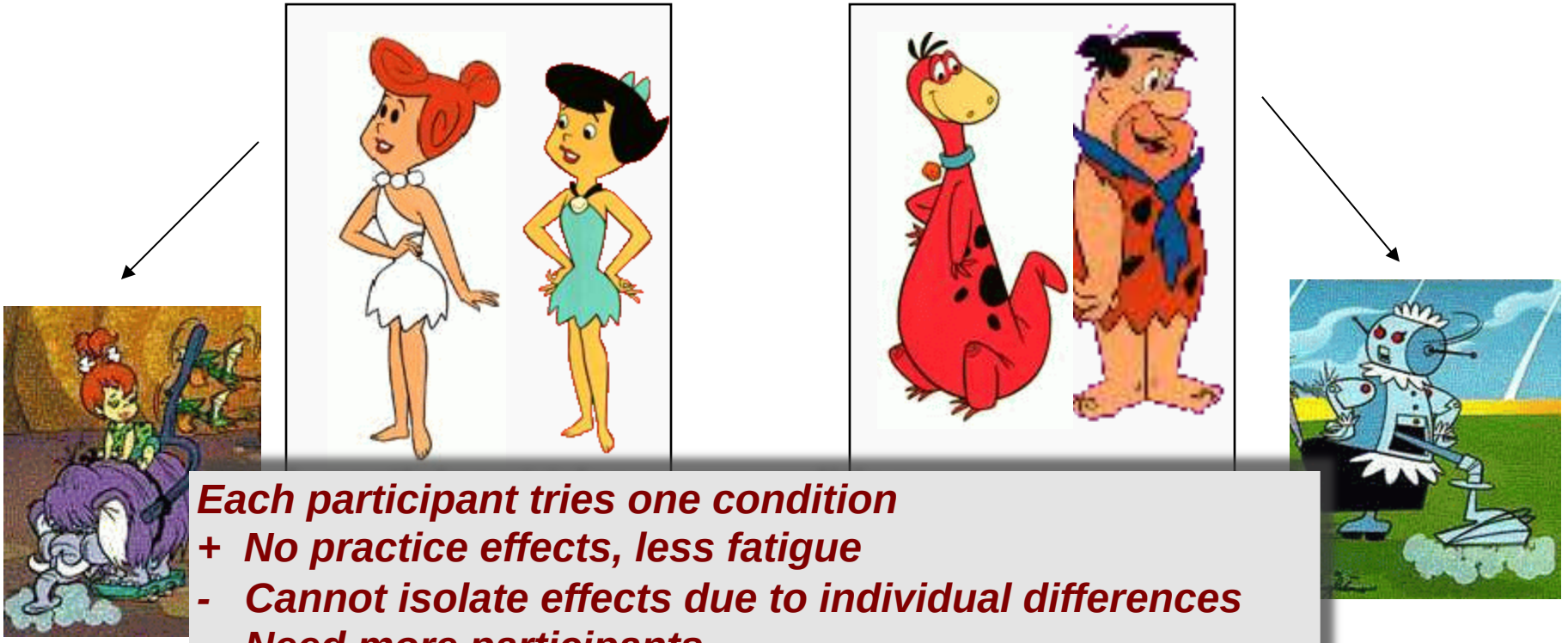
- Assessment of the magnitude of the effect
- No way of telling whether the effect would have occured without the treatment

3. Static Group Comparison Design

Between Groups/ Independent Measures Design

Wilma & Betty use one UI

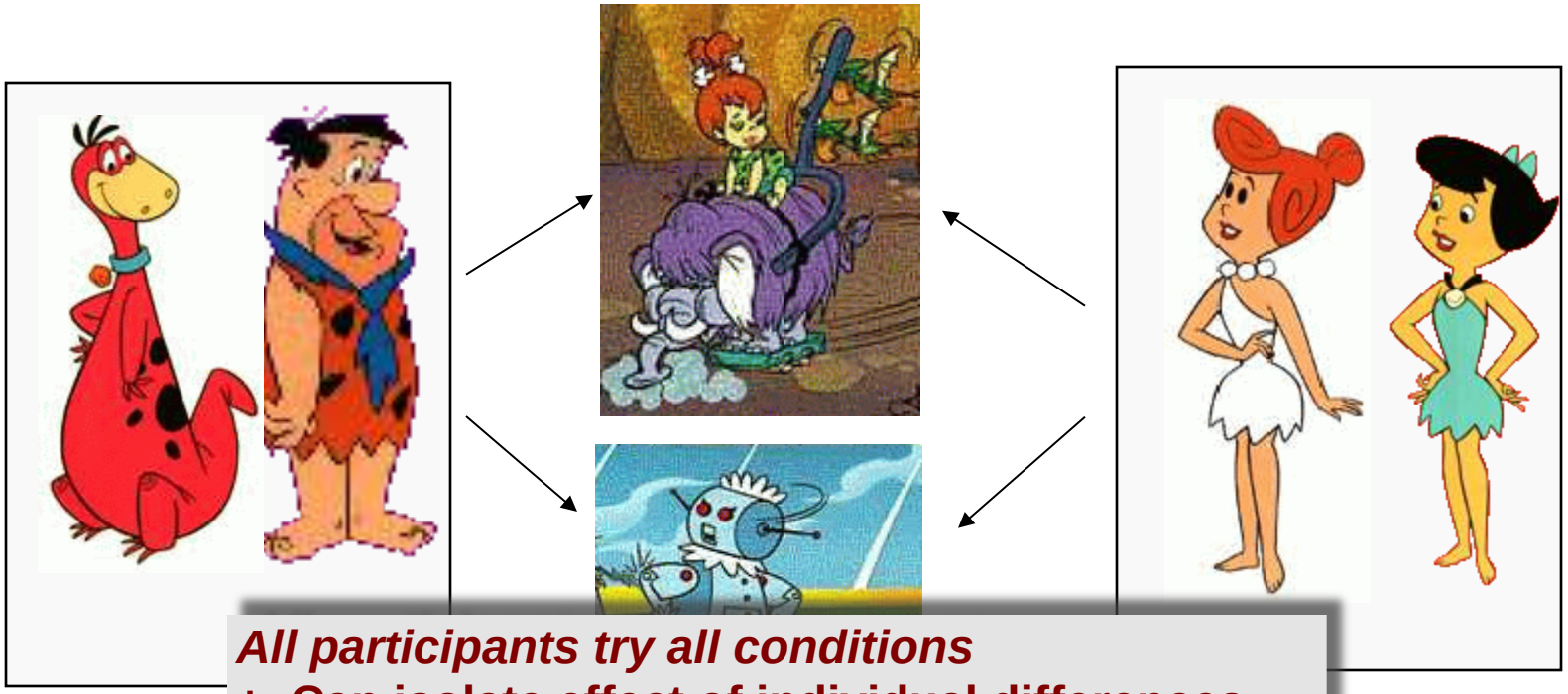
Dino & Fred use another UI



Each participant tries one condition
+ No practice effects, less fatigue
- Cannot isolate effects due to individual differences
- Need more participants

Within Subjects/ Repeated Measures Design

Everyone uses both interfaces



All participants try all conditions

+ Can isolate effect of individual differences

+ Requires fewer participants

- Practice and fatigue effects

Subject (participant) groups in experiments

1. Within subjects („repeated measures“)
   - Each subjects is exposed to all conditions
   - Randomize the order of conditions to avoid ordering affects
2. Between groups (“independent measures”)
   - Seperate groups of participants for each conditions
   - Careful selection of groups is essential
3. Hybrid (“mixed”) designs
   - Combination of between-group and within-subject variables

	Pros	Cons
Within subjects	• Fewer participants required (n)	• Carry-over (learning) effects • Sometimes impossible (e.g. gender)
Between groups	• No carry-over effects • Less fatigue	• More participants required (n * [number of conditions]) • Usually harder to show significance

The importance of randomization

- In all types of experiments randomization is crucial:
  - In within-subject designs ⇨ order of conditions
  - In between-group designs ⇨ allocation to groups
- If you fail to randomize your results can not be interpreted
- Example (between groups): Milk experiment in the 1930ies
  - Huge and expensive experiment with 20 000 school children
  - Examine neutricial effects of milk
  - Teachers „randomly" assigned children to
    - » Experimental group (received milk every day)
    - » Control group (received no milk)
  - Teachers subconsciously tended to assign poor children to the experimental group
  - Result:
    - » Control group were by far superior in weight and height
    - » The whole study was worthless due to the lack of randomization

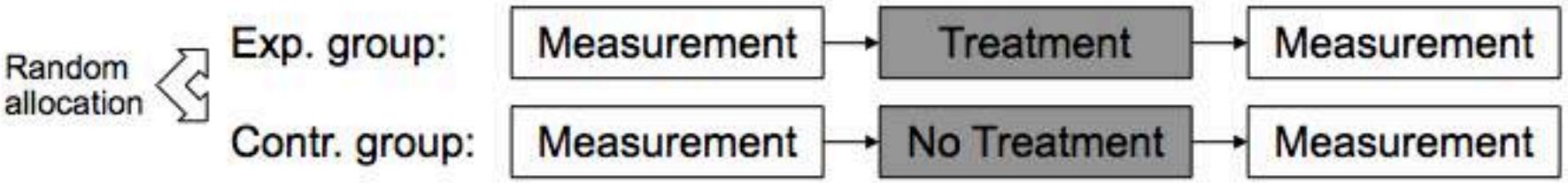
Experimental design: "between groups"

Objective: randomized group allocation ⇨ avoid group threats

1. Post-test only control group design

- Weakness: no way of knowing if randomization fails to produce equivalence

2. Pre-test / post-test control group design



- Pre-test guarantees equivalence
- Weakness: pre-test might affect the performance

Which experimental design should I use?

- Short answer: It depends.
- Long answer:
  - if possible experimental design rather than quasi-experimental
  - Repeated measures is generally easier to set up than between-groups
  - Limit the number of variables to investigate
  - KISS (Keep It Simple Stupid)

Example user study variables

- Imagine you want to compare different mobile phone input methods:
  > T9 vs. Multi-Tab (2 conditions)
- Dependent variables?
  > Time
  > # Errors
- Independent variables?
  > Input method: 2 levels: Multi-tap and T9
  > Text to input: 1 level: text with about 10 words

Example user study hypotheses & experiment design

- Hypotheses

  H-1: Input by multi-tap is quicker than T9

  H-2: fewer errors are made using multi-tap input compared to T9

- Null-Hypotheses

  H0-1: No difference in the input speed between multi-tap and T9

  H0-2: No difference in the number of errors between multi-tap input and T9

- Experimental Method

  > Within subjects

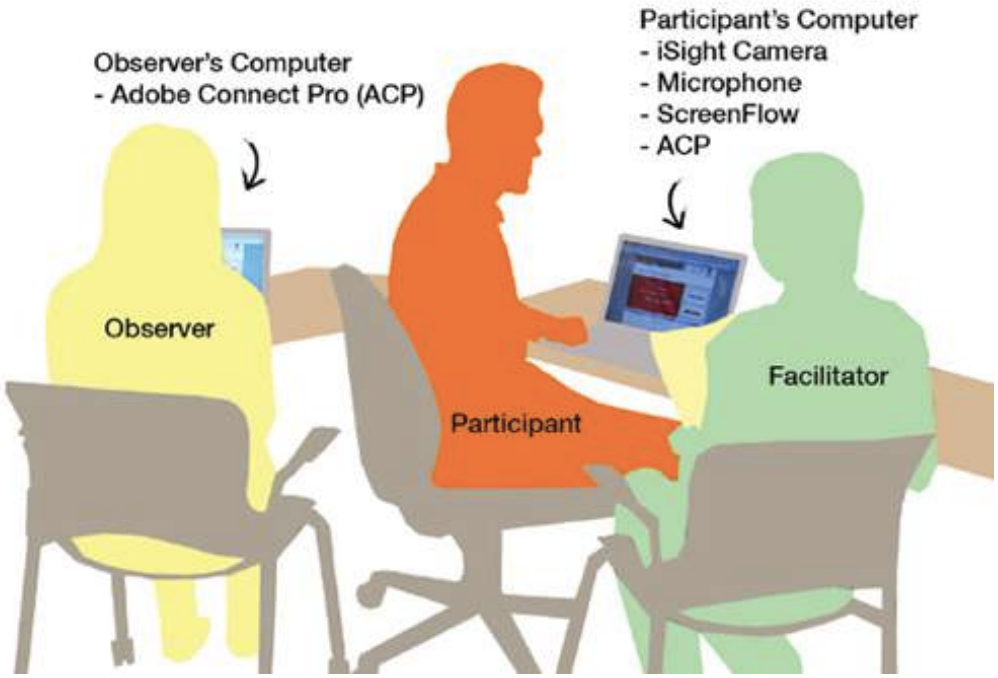
  > Randomized order of conditions

  > Users 1, 3, 5, 7, 9 and 11 perform T9 then Multi-tap

  > Users 2, 4, 6, 8, 10 and 12 perform Multi-tap then T9

Example user study experiment design issues

- Different texts in first and second run?
  - Variable "text" would have two levels
    - ⇨ 4 experimental conditions:
      - » Users 1, 5 and 9 perform T9/Text1 then Multi-tab/Text2
      - » Users 3, 7 and 11 perform T9/Text2 then Multi-tab /Text1
      - » Users 2, 6 and 10 perform Multi-tab/Text1 then T9/Text2
      - » Users 4, 8 and 12 perform Multi-tab/Text2 then T9/Text1
- Particular phone model?
- How to measure
  - Completion time (e.g. stop watch or application?)
  - Number of errors/corrections observed
- Participants
  - How many?
  - Skills
  - Computer user, Phone/T9 users?



PRACTICAL ASPECTS OF STUDY DESIGN

Recruitment of participants

- The number of subjects needed depends on
  - Project
  - Goals
  - Setup

  Minimal size is about 10 subjects
- Participants should be representative for the user group
  - Age
  - Background (e.g. technical vs. not technical)
  - Skills
  - Experience
  - …

  In most cases your team members are NOT representative!

Specification of the experiment setup

The experiment should be set up to be reproducible

⇨ write a specification describing everything which is necessary for reproducing the experiment:

- Hard- and software in use
- Detailed description of self-build prototypes
- The environmental conditions
  - Light conditions
  - Atmosphere
  - ...
- Skills of the test users, e.g.
  - „All participants have to be professional designers"
  - "The candidates should have no experience on using eye-trackers"
  -

Reporting the results

- Anonymize participants
- Background of participants
- Details of tasks, exact wording
- What did they do?
- Why did they do it?
- What didn't they do?
- What is interesting?
- What was surprising to you?

Setting Goals: Developing Test Plan

Scope
- What are you testing?

Purpose
- What concerns, questions, and goals is the test focusing on?

–

Participants
- How many users of what types will you recruit?

Scenarios
- What will participants do with the product in this round of testing?
- What roles will be tested?

Questions
- What will you ask at the beginning and end of the session?

Data to be collected
- What will you count?

Data collection strategy
- make sure your data really will answer your questions

Set up
- What system will you use for testing?
- Will you be videotaping and/or audio-taping?
- Will you be using a specific technology to capture data?

A Good Usability Test Plan …

- a goal/task
  - what to do
  - what question to find the answer for
- data to be used in the task
- elaborated scenario based on the initial goal/task
- screening questionnaire for participants

Selecting/Preparing Test Tasks

- Should reflect real tasks (and their order)
- Choose one simple order (simple → complex)
  - tasks from analysis & design can be used
- Avoid bending tasks in direction of what your design best supports
- Don't choose tasks that are too fragmented
- Provide training if needed
- Assign enough time to finish the task (to ensure all finish)

Exploratory vs. Evaluative Questions?

- *exploratory* = early design "*test drive*," problem identification (usually qualitative)
- *evaluative* = "hit the mark," meet the interim or release criteria (usually quantitative)
- what do you *need* at this point?
- combinations can be very effective

Questions

Are icons intuitive in our application?

Exploratory questions:

- When we specify an action in a task, do users choose the right icon?
- What happens when they try?
- What problems do they have?

Evaluative question:

- When we specify an action in a task, do users choose the right icon on *the first try at least 80% of the time*?

Should our design include a tool bar?

Evaluative question:

- What is the general advantage of a tool bar?
- Which specific actions the tool bar supports (e.g., visible palette)?
- Does it improve the efficiency (e.g. one-click away)?

Deciding on the Data to Collect

- process data
  - observations of what users are doing & thinking
  - "thinking aloud"
    - what they are thinking
    - what they are trying to do
    - questions that arise as they work
    - things they read
- bottom-line data
  - summary of what happened (time, errors, success)
  - i.e., the dependent variables

Measuring Bottom-line Usability

- Situations in which numbers are useful
  - time requirements for task completion
  - successful task completion
  - compare two designs on speed or # of errors
- Ease of measurement
  - time is easy to record
  - error or successful completion is harder
    - define in advance what these mean
- Do not combine with thinking-aloud. Why?
  - talking can affect speed & accuracy

Measuring User Preferences

- How much users like or dislike the system
  - can ask them to rate on a scale of [1..5], [1..7] [1..10]
  - or have them choose among statements
    - "best UI I've ever…", "better than average"…
  - hard to be sure what data will mean
    - novelty of UI, feelings, not realistic setting …
- If many give you low ratings → trouble

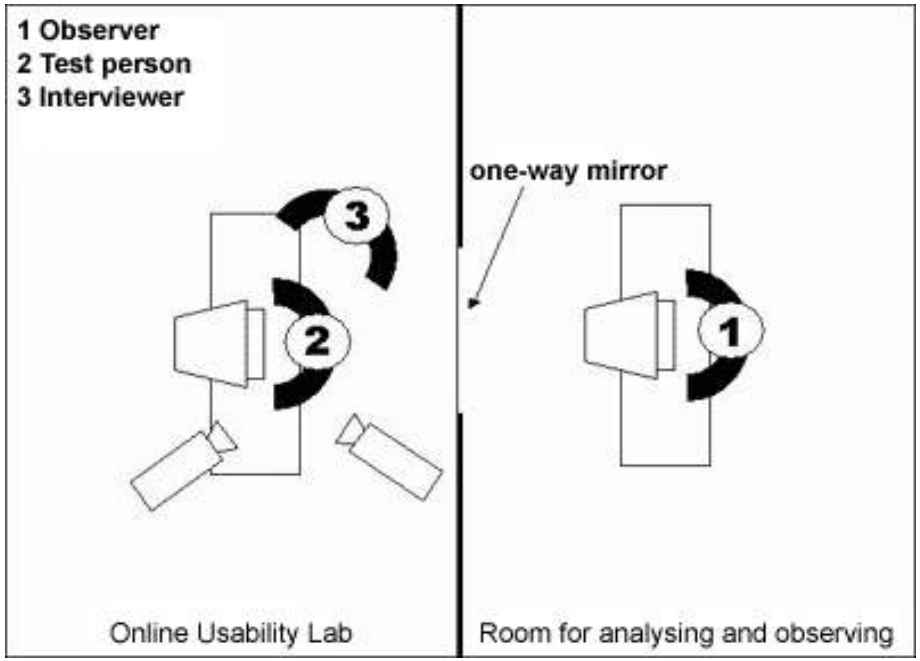
- Can get some useful data by asking
  - what they liked / disliked
  - where they had trouble, best part, worst part
    - redundant questions are OK

Test Session

- Number of prototypes matching the number of participants
- Monitoring, observation facilities (e.g. for user logs, screen capture video, audio comments, hand-written notes)
- Consent forms for participants to sign
- Testing procedure description & testing scenarios
  - Pre-test
  - Test
  - Post-test
- Do a dry-run and a pilot test
  - helps you fix problems with the study
  - do 2, first with colleagues, then with real users

Usability Testing

Usability Testing @ VU MexiaXperience



You can reserve the testing room @ VU MediaXperience for your sessions

http://www.vumediaxperience.nl/html/videorefl.html

Usability Testing @ VU IntertainLab

You can reserve the testing room @ VU IntertainLab for your sessions

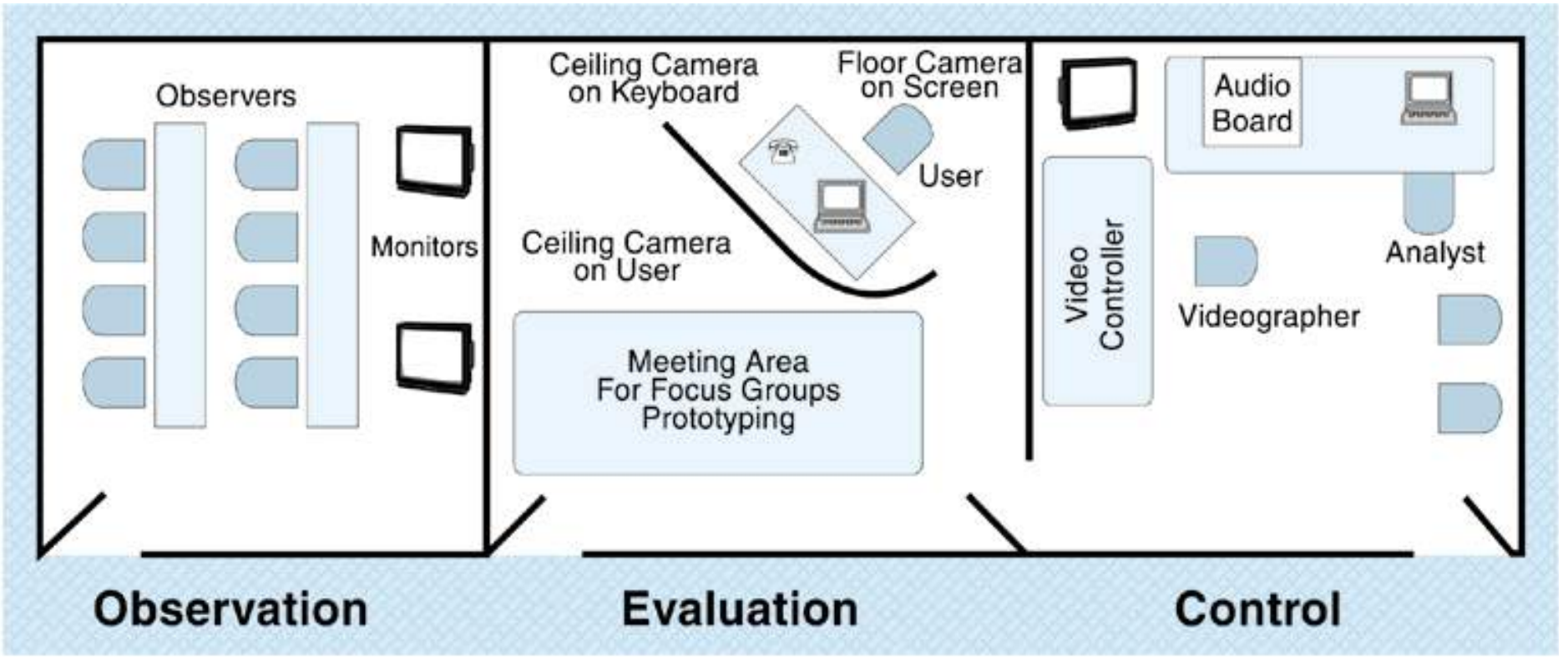
http://www.networkinstitute.org/tech-labs/intertain-lab/

Usability Testing @ VU MediaLab

You can reserve the testing room @ VU MediaLab for your sessions

http://www.networkinstitute.org/tech-labs/medialab/

Usability Test Setting



IN SUMMARY

Recap

- Usability testing is necessary to assess ***how well an interface/system/app works***
- Evaluation experiments should aim for ***reliability, validity and generalisability***
- ***Different experimental setups possible***: between groups, within groups, single subject and single vs. multiple dependent variables
- Experimental design depends on ***available users, variables to be tested and goal of study***

Acknowledgements

Part of the slides are from Sara Streng,
Ludwig-Maximilians-Universität, Munich