Mensch-Maschine-Interaktion

User Study Design

Sara Streng
Ludwig-Maximilians-Universität München
Sommersemester 2009

Agenda

2. User Study Design

- 2.1. The Purpose of User Studies
- 2.2. Research Aims: Reliability, Validity and Generalizability
- 2.3. Research Methods and Experimental Designs
- 2.4. Ethical Considerations
- 2.5. HCI-related and practical information for your own studies

Agenda

2. User Study Design
    - 2.1. The Purpose of User Studies
    - 2.2. Research Aims: Reliability, Validity and Generalizability
    - 2.3. Research Methods and Experimental Designs
    - 2.4. Ethical Considerations
    - 2.5. HCI-related and practical information for your own studies

The Purpose of User Studies

What are user studies needed for?

- “To learn more”
- To ensure quality in product development
- To compare solutions
- To provide quantitative figures
- To get a scientific statement (instead of personal opinion)

Examples of scientific statements

- 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

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

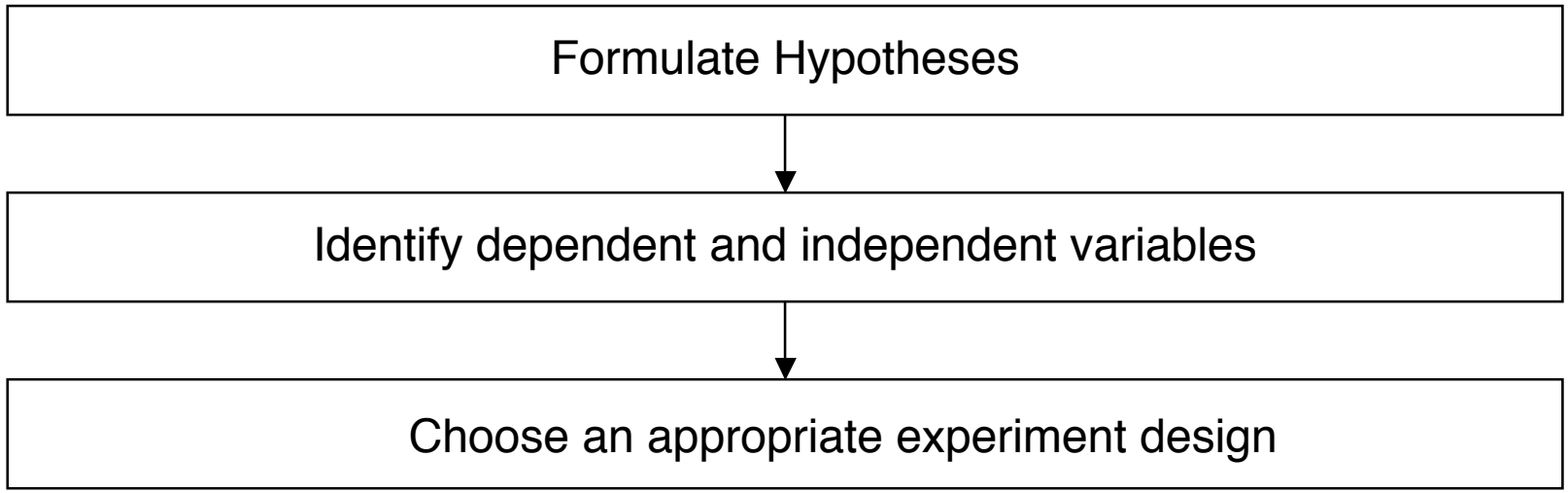
- Only way to infer causility:
  - Two controlled situations
    1. Cause is present (*experimental condition*)
    2. Cause is absend (*control condition*)
  - Otherwise the situations have to be identical!

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)

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!



How to Isolate the Cause

1. Control conditions
2. Controlling other factors

   ⇨ Minimize the risk of other factors influencing the experiment
3. Randomizing allocation of participants to experimental and control group

   Example: Instruction Manual
   - RQ: Does reading a manual help to use a device (e.g. a mobile phone) more efficiently?
   - Conditions:
     1. Experimental condition: Participants read the manual
     2. Control condition: Participants do not read the manual
   - About half the participants own this device. Imagine all of them would be allocated to the experimental condition, the other ones to the control condition. What happens?

Agenda

2. User Study Design
    - 2.1. The Purpose of User Studies
    - 2.2. Research Aims: Reliability, Validity and Generalizability
    - 2.3. Research Methods and Experimental Designs
    - 2.4. Ethical Considerations
    - 2.5. HCI-related and practical information for your own studies

Aims of Research

The results of your experiment should be

1. Valid
   - ⇨ Results should be accurate
   - ⇨ Results should show what you intend to show
2. Reliable
   - ⇨ Results should be potentially replicable by anyone
3. Generalizable
   - ⇨ Results should have a wider application than the particular circumstances of the experiment
4. Important

In order to be (potentially) important the results need to fulfill the first three criteria!

Reliability

- Consistency of measurement: Degree to which an instrument measures the same way each time it is used under the same condition with the same subjects
- A measure is reliable if a person‘s score on the same test given twice is similar.
- Two ways of estimating reliability:
  1. Test/Retest
     - Conservative method
     - Two separate times of measurement
     - Compute correlation between the two measurements
     - Assuming the conditions are the same
  2. Internal Consistency
     - Group questionnaire items that measure the same concept
       e.g. two sets of questions that both measure motivation
     - Compute correlation between the two sets
     - **Cronbach‘s Alpha**: split all questions every possible way and compute correlations for all of them ⇨ correlation coefficient

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

Example: Brain Weight

- Paul Broca investigated human abilities / intelligence by measuring brain weight (19th century)
- Findings:
  - Brain of Caucasian men > Brain of Caucasian women > Brain of negroes
  - Brain of French men > Brain of German men
- Is brain weight a true score for intelligence?
  - No, because it is known that within all species there is no relationship between brain weight and intelligence
- What other things does brain weight reflect?
  - Relation to body size
  - Age (mainly elderly females and young males, who died in car accidents)

Reliable? ☑

Valid? ☒

Example: Folding Rule

- A folding rule is only 1.9 m instead of 2 m
- Everytime it is used to determine the length of an object, it systematically overestimates the length.

Reliable? ☑

Valid? ☒

Threats to Internal Validity (1)

- Group threats
  - If experimental and control group are different the study is worthless
- Instrument change, e.g.
  - Different measuring devices
  - Interviewer gets more practised
- Reactivity and experimenter effects
  - Measuring a person‘s behavior might already change the behavior
  - Social desirability
  - Ideally: Double-blind technique (participant and experimenter unaware of hypotheses and conditions)
- Differential Mortality
  - When testing the same individuals repeatedly
  - E.g. pre-test is not comparable to post-test when many participants drop out

Threats to Internal Validity (2)

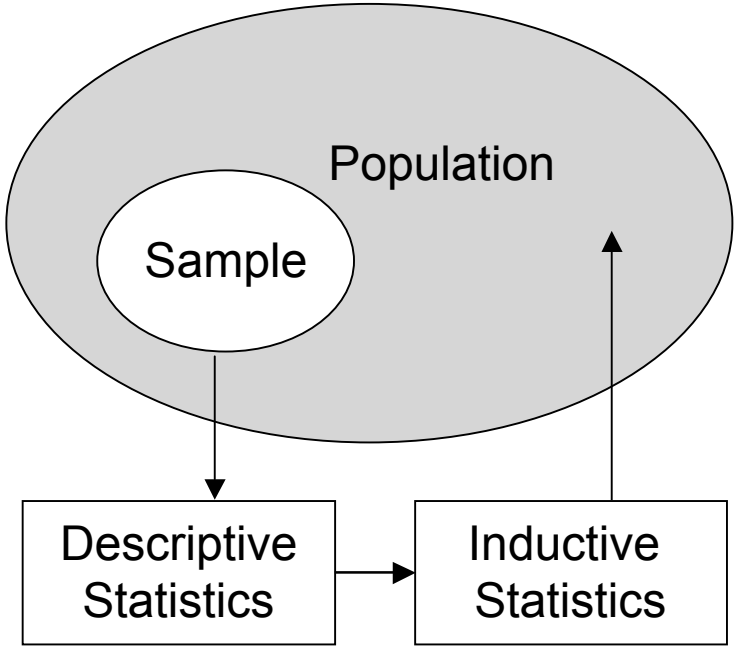
- Regression to the mean
  - If extrem scores were produced on a pre-test, it is more likely that the score is closer to the mean on a subsequent test
  - Problem always occurs when measuring the effect of a problem solution / policy
- Time threats
  - Maturation, e.g. children‘s reading ability
  - Influence of events unrelated to the manipulation that occured during the treatment, i.e. between pre- and post-test

Threats to External Validity

- Over-use of special participant groups
  - McNemar 1946: „psychology is largely a study of undergraduate behavior“
  - 70-90% of participants are undergraduates (Rosenthal and Rosnow, 1975)
  - Today: how valid are experiments that are done with Media Informatics students only?
- Restricted numbers of participants
  - Typical threat to reliability
  - Also affects the ability to generalize

Generalizability

- What do we want to gain from a user study?
  - Result, which is valid for all people
- Test users must be representative
- Descriptive statistics:
  - Tables
  - Diagrams
  - Means
  - ...
- Inductive statistics:
  - Ensure validity for the whole



Quality of Study Design

- Well designed experiments isolate causal factors well
- Poor designed experiments leave many alternative explanations of the results ⇨ practically useless
- Data consists of four components:
  1. A „true score“ for the things we hope to measure — maximize
  2. A „score for other things“ that are measured inadvertently — minimize
  3. Systematic (non-random) bias — minimize
     - Should (if at all) affect all participants in the study
  4. Random (non-systematic) error — minimize
     - Should be cancelled out over large numbers of observations

Agenda

2. User Study Design
    - 2.1. The Purpose of User Studies
    - 2.2. Research Aims: Reliability, Validity and Generalizability
    - 2.3. Research Methods and Experimental Designs
    - 2.4. Ethical Considerations
    - 2.5. HCI-related and practical information for your own studies

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

Compromise:

- Varify causal hypotheses ⇨ confirm findings with more natural observations or
- Identity hypotheses through observations ⇨ varify hypotheses in experiments

Quasi-Experimental Method

1. Observational
   - No manipulation
   - Record behavior systematically and objectively
   - Strength: observe people how they behave normally (e.g. driving behavior)
   - Downside:
     - No identification of cause and effect
     - Time consuming
2. Quasi-experimenal
   - Sometimes real experiments are not possible (e.g. for ethical reasons)
   - Control over timing of measurement
   - No (complete) control over independent variables
     ⇨ Impossible to isolate cause and effect
3. Experiment
   - Manipulation by experimenter
   - Only way to prove cause and effect

Quasi-Experiment - Example: Motorcyclists

- 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
    - » ...

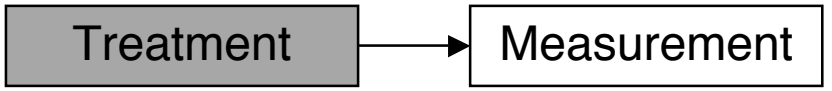
Experiments on Age- and Gender-Differences

- E.g. is there an age-difference in problem-solving ability?
- Most researchers investigate effects of age and gender as „true“ experiments
- Strictly speaking, they are quasi-experiments:
  - Participants are not randomly allocated to the groups
  - Impossible to rule out other reasons than age or gender difference, such as
    - » Born at different times
    - » Different life experience
    - » ...

⇨ Be aware of the complications in interpreting the results!

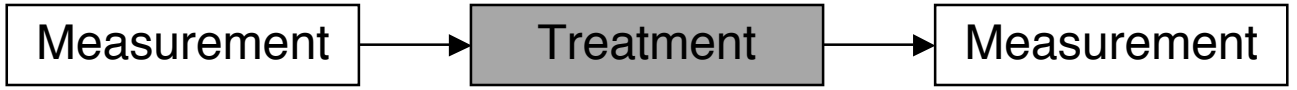
Types of Quasi-Experimental Designs (1)

1. One group post-test design



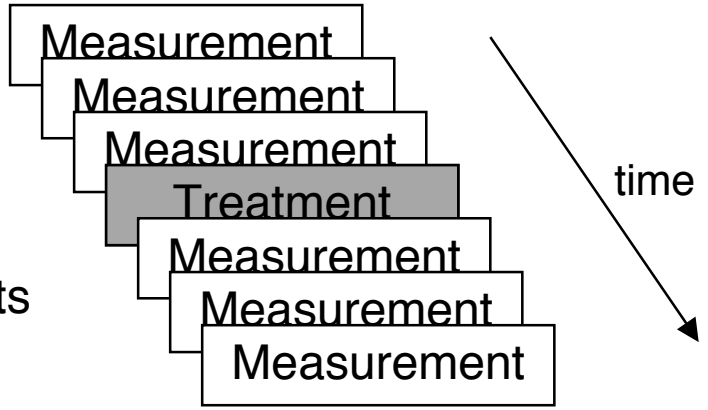
   - No baseline against which to compare

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



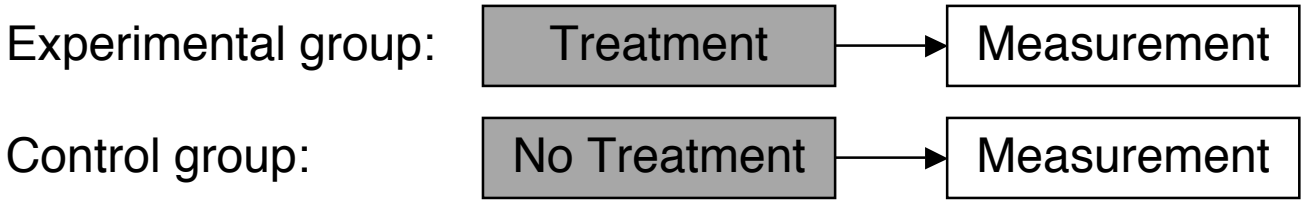
   - Assessment of the magnitude of the effect
   - No way of telling whether the effect would have occured without the treatment

3. Interrupted time-series design
   - Series of measurements, some before some after the treatment
   - Weakness: Not immune to history threats



Types of Quasi-Experimental Designs (2)

4. Static Group Comparison Design



- Experimental group with treatment
- Control group without treatment
- Only weakness: Participants are not randomly assigned
- See motorcyclist example

Types of Experimental Designs

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

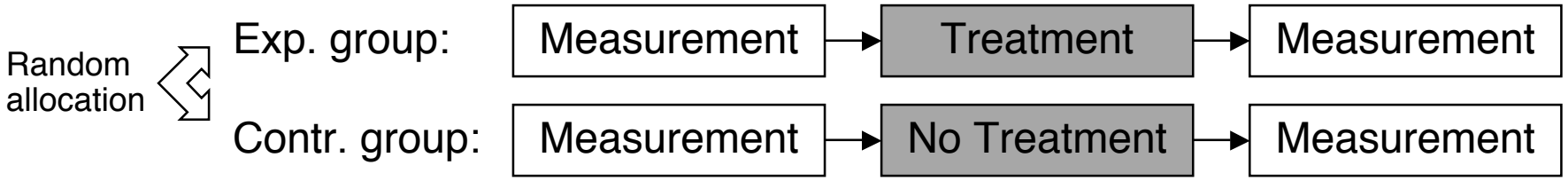
Types of Between Group Designs (1)

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

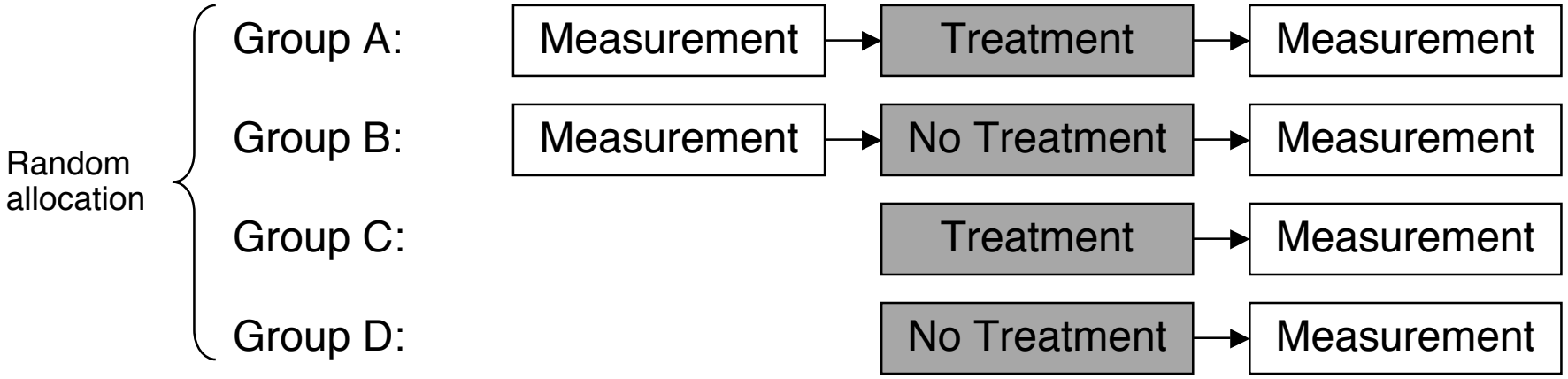
2. Pre-test / post-test control group design



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

Types of Between Group Designs (2)

3. Salamon four-group design



- Two experimental groups (A and C)
- Two control groups (B and D)
- Groups A and B show the effects of presence/absence of the treatment
- Groups C and D show the effects of the pre-testing
- Very expensive in time and # participants

  ⇨ rarely used

Types of Within Subject Designs

- Objective: random / counterbalanced order of conditions
- Trivial for 2 conditions:
  Half of the participants start with condition A, the other half with condition B
- For more than 2 conditions:
  - Randomize order
  - Systematically counterbalance the order (Latin Square Design):
    - » There are n! different orders for n conditions
    - » Instead of running n! different orders (= groups), only use n and still avoid order effects
    - » Idea: Every condition is placed at each „position“ once
    - » Each order is used by one of the n groups of participants
    - » Weakness:
      - Unbalanced for odd numbers of conditions
        e.g. n = 3: A before B twice, B before A once

n = 3 ⇨ ABC, BCA, CAB
n = 4 ⇨ ABCD, BADC, CDAB, DCBA

Multi-Factorial Designs

- All designs covered so far: manipulation of only 1 variable

  Note: Do not confuse
  - Multiple levels of one variable (e.g. different doses of a drug) with
  - Multiple variables (e.g. (1) different drugs taken at (2) different times of the day)

- Advantage of using multiple variables:
  - Analyze how multiple variables interact
  - Not much extra work in within subject designs (only more task(s))
- Disadvantage:
  - Experiments with more than 2 - 3 variables are difficult to interpret!
  - Much extra work in between group designs (#groups multiplies)
- Number of experimental conditions = product of the variables' levels, e.g.
  - Font can be *Arial* or *Times* (2 levels)
  - Background can be blue, green, or white (3 levels)

  ⇨ 6 experimental conditions

Agenda

2. User Study Design
    - 2.1. The Purpose of User Studies
    - 2.2. Research Aims: Reliability, Validity and Generalizability
    - 2.3. Research Methods and Experimental Designs
    - 2.4. Ethical Considerations
    - 2.5. HCI-related and practical information for your own studies

Ethical Considerations

- Be aware of the influence and power of the experimenter
- Responsibility to the participants!
- Some research institutions have an ethics committee, which examine details of your proposed study before you can run the experiment.
- If not, you should still follow some guidelines:
  - Protect the participants‘ confidentiality
  - Protection from physical and psychological risks (of psychological or medical experiments)
  - Informed Consensus: Inform participants about:
    - The experiment (in particular risks)
    - Their rights (in particular withdrawal from the study)
    - Confidentiality
  - Inform participants, that the system is evaluated - not the user.
    - If something does not work, it is never the user‘s fault!
  - Debriefing: Tell participants what the study was about in the end

Procedure

1. Set goals (hypotheses)
2. Design the experiment
3. Do a pilot study
4. Recruite users
5. For each user, typically:
   - Inform the user about the experiment (see next slide)
   - Consent form
   - Do a survey on
     - » Demographics
     - » Questions related to the experiment (e.g. left- / right-handedness)
   - Give instructions on the task
   - Let the user do the tasks and measure the variables
   - Be available for questions and (informal) feedback
6. Analyze the results ⇨ accept / reject hypotheses

Informing the Participants About the Study

Inform the participant about:

- General purpose of the study
- Procedure
  - Amount of time
  - Breaks
  - ...
- Their right to withdraw from the study at any time
- Confidentiality
- Risks
- The system is evaluated - not the user:
  Interest is in aggregated data of all participants, not in the individual ones!

Never reveal:

- Hypotheses
- Conditions

Consent Form

- **Participants Consent Form**

- **Study** _____ **Institution** _____

- Name: _____ Date of Birth: _____
- Email: _____
- Phone:_____

- I have been informed on the procedure and purpose of the study and my questions have been answered to my satisfaction.
- I have volunteered to take part in this study and agree that during the study information is recorded (audio and video as well as my interaction with the system). This information may only be used for research and teaching purpose. I understand that my participation in this study is confidential. All personal information and individual results will not be
- released to third parties without my written consent.

- I understand that I can withdraw from participation in the study at any time.

- Date: _____ Signature:_____

Agenda

2. User Study Design
    - 2.1. The Purpose of User Studies
    - 2.2. Research Aims: Reliability, Validity and Generalizability
    - 2.3. Research Methods and Experimental Designs
    - 2.4. Ethical Considerations
    - 2.5. HCI-related and practical information for your own studies

What is Evaluated in HCI Research?

- 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

Qualitative vs. Quantitative User Studies

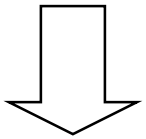
Qualitative:
- Get “non-measurable” feedback
- General insight

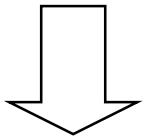
- Used to
  - Find problem areas
  - Find conceptual errors
  - Find missing functionality

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

Recruiting and 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 Results of a User Study

- 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?

(based on
https://apps.lis.uiuc.edu/wiki/download/attachments/2654987/User+study.ppt)

What You Should Keep in Mind

- Don’t learn how to conduct the experiment during the user study.
  Think about what to do in case of problems in advance,
  e.g. how to proceed if the mobile phone of a user gets an incoming call during a test run?
  - Stop the recording and repeat the test run?
  - Stop the test and don’t use the data?
- Times can be recorded automatically by the testing software or stopped manually with a watch.

Example User Study Design - 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 Design - Hypotheses

- 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 Design– Other Aspects

- 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?