universitätbonn

Useable Security and Privacy

Usability & Evaluation

Thanks go to:
- Prof. Dr. Michael Rohs

universitätbonn

USABILITY

User – Tool – Task/Goal – Context

context

user

tool

task/goals

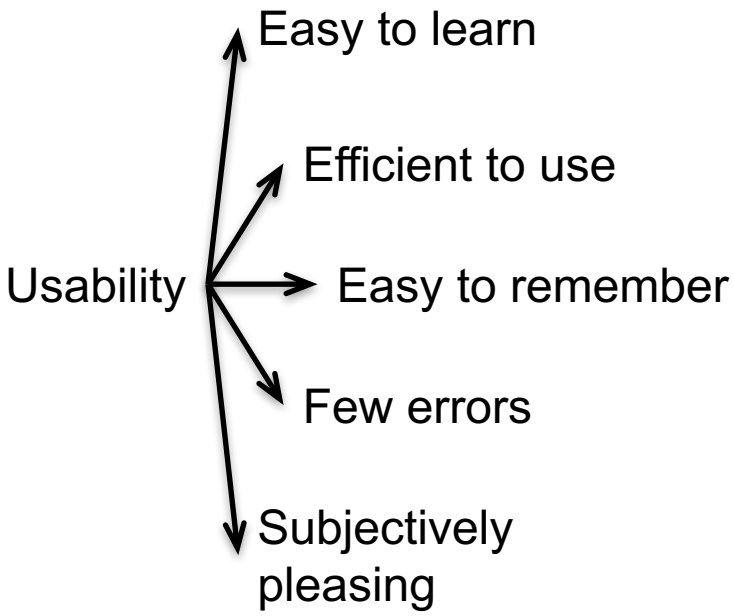
Usability (ISO 9241 Standard)

- Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.
  - Effectiveness: Quality, accuracy, and completeness with which users achieve goals
  - Efficiency: Effort necessary to reach a certain level of quality, accuracy, and completeness
  - Satisfaction: Comfort and acceptability of the system to its users (enjoyable, motivating? or limiting, irritating?)
  - Context of use: Users, tasks, equipment, physical and social environment, organizational requirements

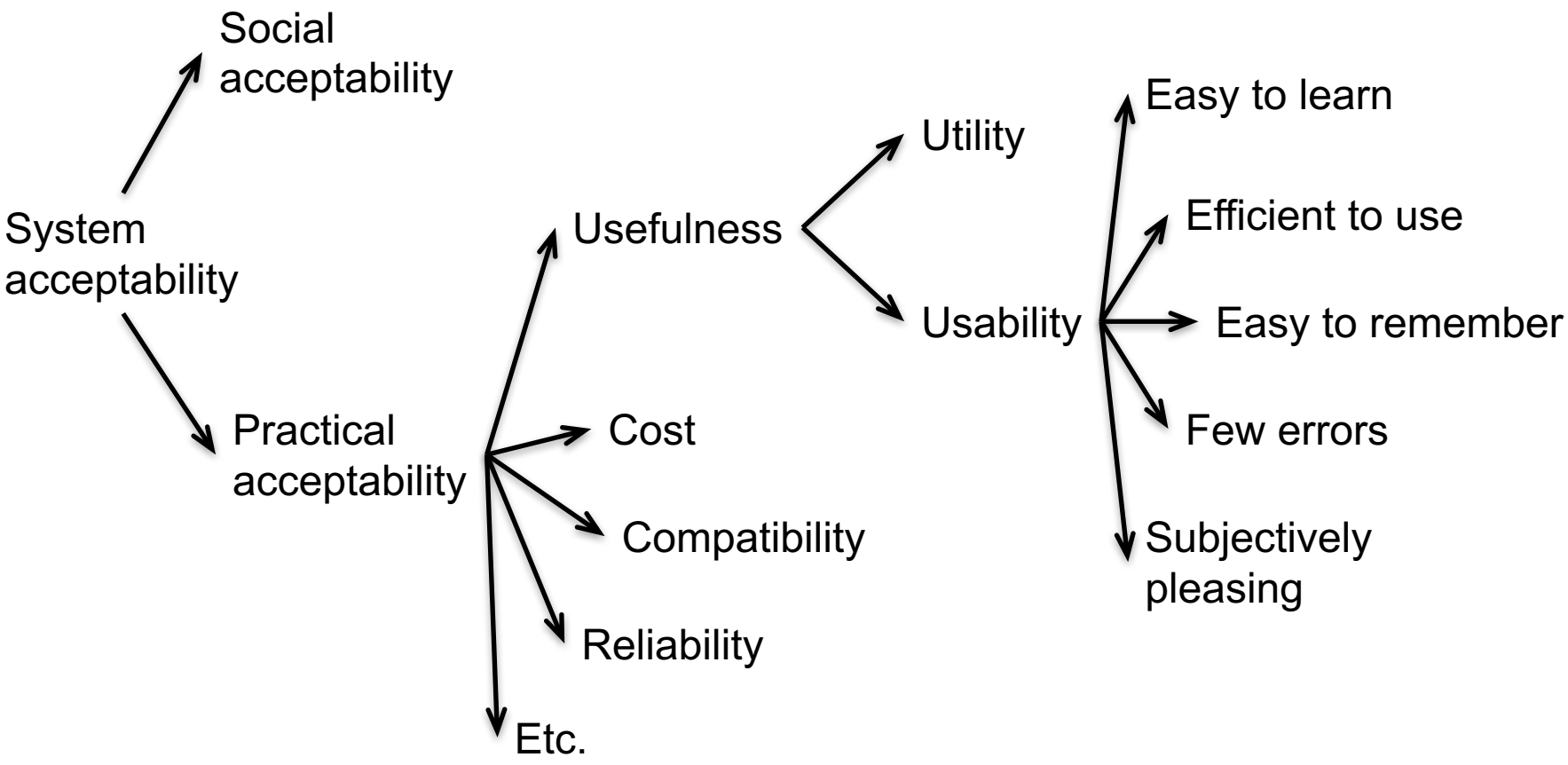
ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11: Guidance on usability—Part 11 (ISO 9241-11:1998)

Attributes of Usability (Nielsen)

- Learnability (easy to learn)
- Efficiency (efficient to use)
- Memorability (easy to remember)
- Errors (few errors)
- Satisfaction (subjectively pleasing)



Usability as an Aspect of System Acceptability (Nielsen)



Typical Measures of Effectiveness

- Binary task completion
- Accuracy
  - Error rates
  - Spatial accuracy
  - Precision
- Recall
- Completeness
- Quality of outcome
  - Understanding
  - Experts' assessment
  - Users' assessment

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. Int. J. Human-Computer Studies 64 (2006) 79–102.

Typical Measures of Efficiency

- Time
  - Task completion time
  - Time in mode (e.g., time in help)
  - Time until event (e.g., time to react to warning)
- Input rate (e.g., words per minute, WPM)
- Mental effort (NASA Task Load Index)
  - http://www.keithv.com/software/nasatlx/
- Usage patterns
  - Use frequency (e.g., number of button clicks)
  - Information accessed (e.g., number of Web pages visited)
  - Deviation from optimal solution (e.g. path length)
- Learning (e.g., shorter task time over sessions)

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. Int. J. Human-Computer Studies 64 (2006) 79–102.

Typical Measures of Satisfaction

- Standard questionnaires (e.g., SUS, QUIS, AttrakDiff)
- Preference
  - Rate or rank interfaces
  - Behavior in interaction (e.g., observe what users choose)
- Satisfaction with the interface
  - Ease-of-use (e.g. 5-/7-point Likert scale: “X was easy to use”)
  - Satisfaction with specific features
  - During use (e.g., heart period variability, reflex responses)
- Attitudes and perceptions
  - Attitudes towards others (e.g., “I felt connected to X when using…”)
  - Perception of outcome / interaction

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. Int. J. Human-Computer Studies 64 (2006) 79–102.

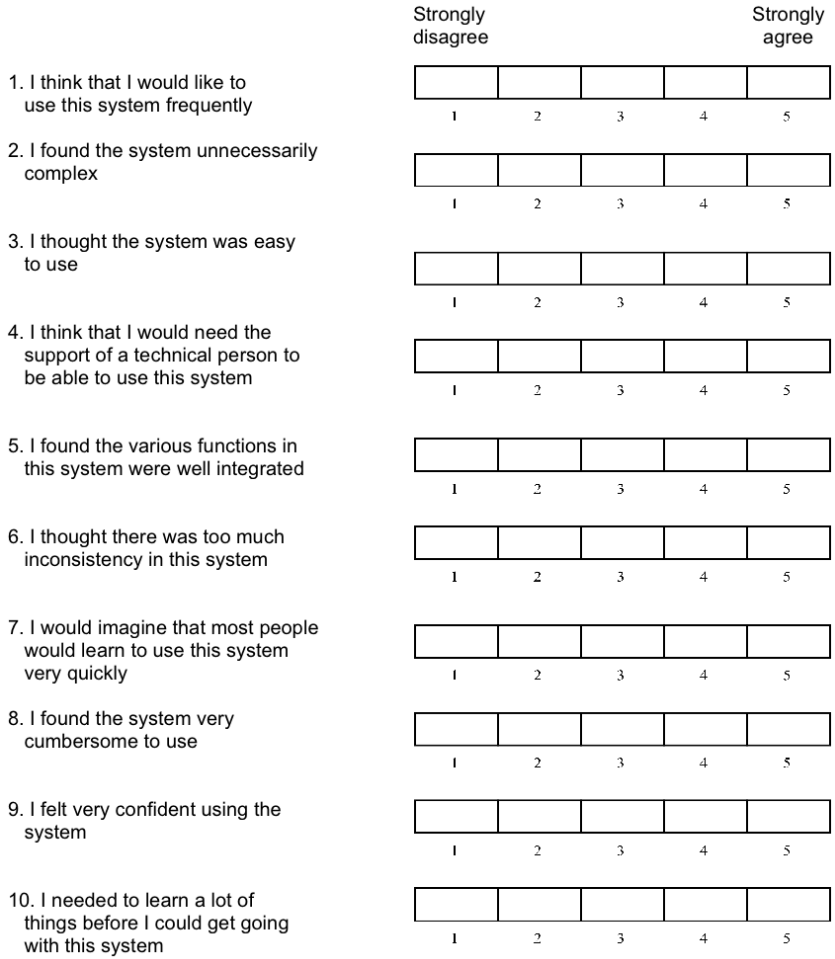
Typical Measures of Specific Attitudes

- Annoyance
- Anxiety
- Complexity
- Control
- Engagement
- Flexibility
- Fun
- Liking
- Wanting to use again

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. Int. J. Human-Computer Studies 64 (2006) 79–102.

SUS: System Usability Scale

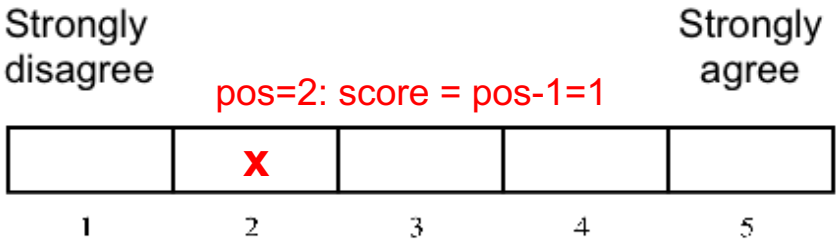
- Developed by DEC Corporation
- 10 5-point Likert scales
- Single score (0-100)
  - Odd items: position – 1
  - Even items: 5 – position
  - Add item scores
  - Multiply by 2.5



Brooke. SUS: A "quick and dirty" usability scale. Usability Evaluation in Industry. London: Taylor and Francis, 1996

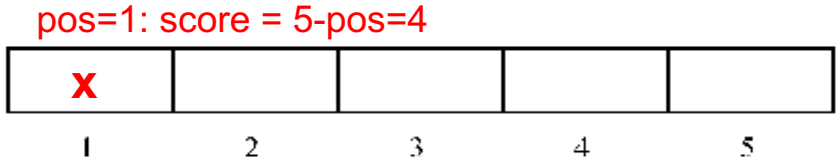
Example: SUS-Ratings

1. I think that I would like to use this system frequently



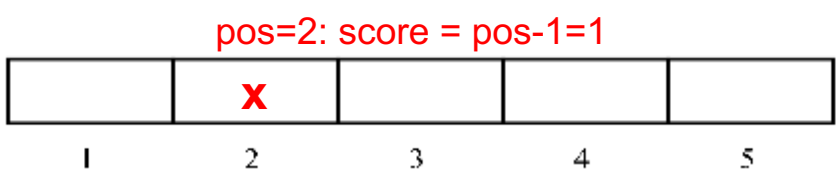
1

2. I found the system unnecessarily complex



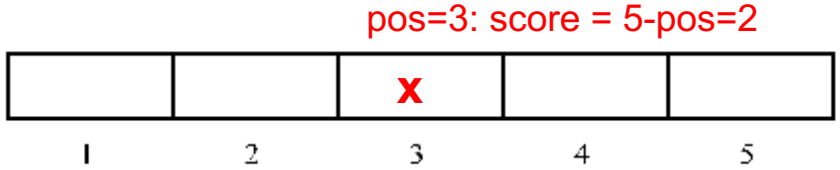
4

3. I thought the system was easy to use



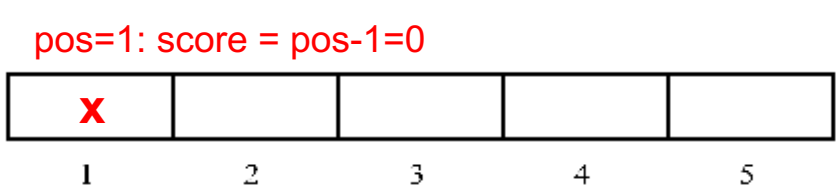
1

4. I think that I would need the support of a technical person to be able to use this system



2

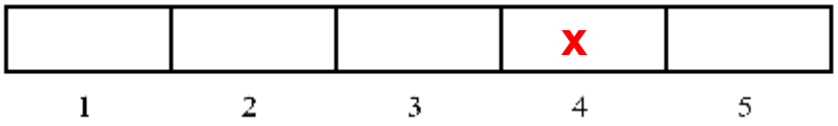
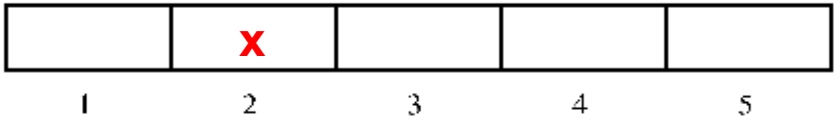
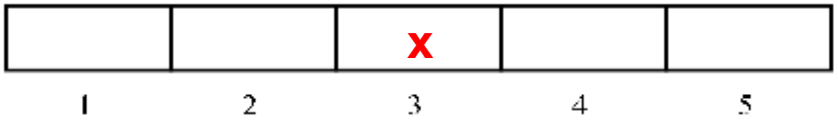
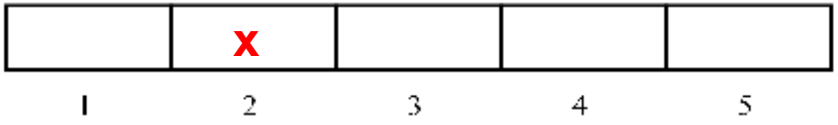
5. I found the various functions in this system were well integrated



0

Brooke. SUS: A "quick and dirty" usability scale. Usability Evaluation in Industry. London: Taylor and Francis, 1996

Example: SUS-Ratings

6. I thought there was too much inconsistency in this system		1
7. I would imagine that most people would learn to use this system very quickly		1
8. I found the system very cumbersome to use		2
9. I felt very confident using the system		1
10. I needed to learn a lot of things before I could get going with this system		3

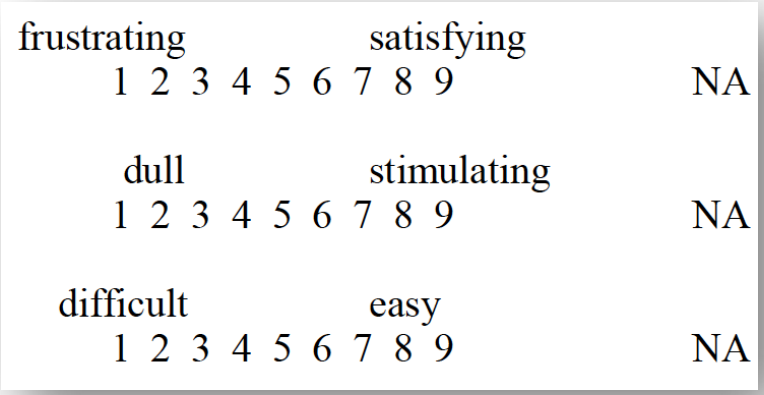
Sum = 16

SUS-Score = Sum * 2.5 = 40

Brooke. SUS: A "quick and dirty" usability scale. Usability Evaluation in Industry. London: Taylor and Francis, 1996

QUIS: Questionnaire for User Interaction Satisfaction

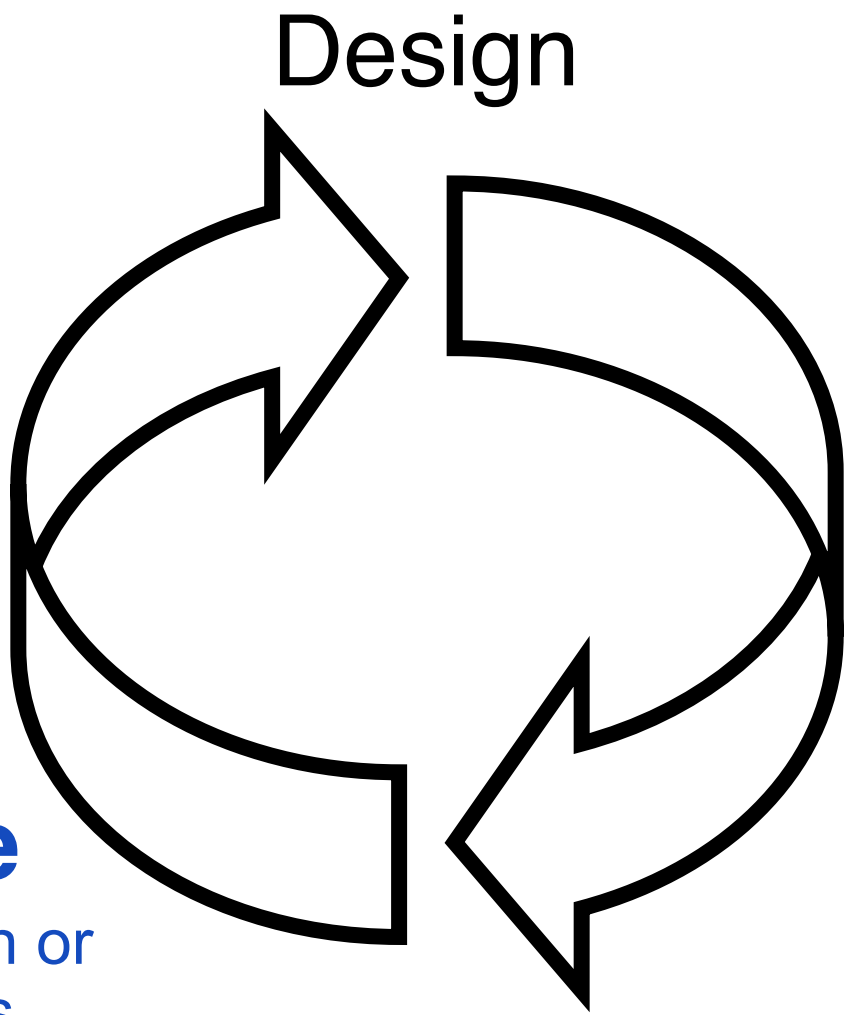
- Developed by the University of Maryland
  - Academic License $200, Student License $50
- Semantic differential scales
- Components: (1) demographics, (2) overall reaction ratings (6 scales), (3) specific interface factors: screen, terminology and system feedback, learning, system capabilities, (4) optional sections
- Long and short forms
- http://lap.umd.edu/quis/



Chin, Diehl, Norman: Development of an instrument measuring user satisfaction of the human-computer interface. CHI '88

EVALUATION

DIA Cycle: When to evaluate?



Analyse

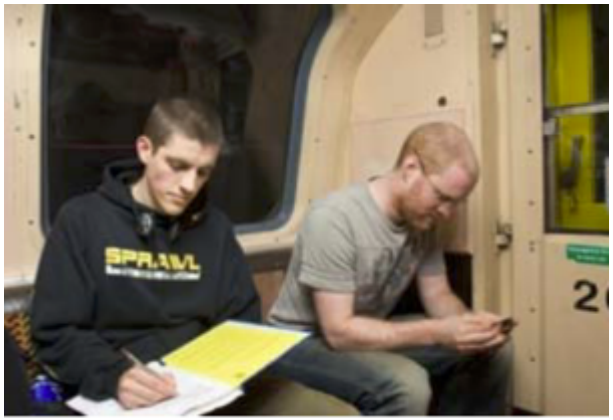
Evaluate with or without users

Where to evaluate: Laboratory

+ Equipment (audio / video, see-through mirrors, special computers), no disruptions, quiet
– Natural environment missing (shelves, wall calendar, streets, people…); unnatural situation (relevance?)

Only place possible if real use un-enthical, remote (ISS…), or controlled situation needed

Where to evaluate: In the field


- Studies in the users' natural environment
- Advantages
  - + Situations (location and context!) and behavior more natural
  - + More realistic (also because of disruptions)
  - + Better suited to long-term studies
- Disadvantages
  - Noise, task interruptions, effort
  - Could still feel like a test situation

Evaluation in the Mobile Context

- Context of use needs to be taken into account
  - Factors: User, activity, device, environment
- Usage “on the move”
  - Physically moving: walking, driving a car, traveling as a passenger
  - Being in different places: away from office environment or home
- Difficult to collect data in the field
  - Recording interaction
  - Capturing context
  - Controlling experimental conditions

Comparison of Lab and Field Tests

- Assess quantity and quality of usability problems found in lab vs. field
- Tasks and scenarios given

Table 1. The Test Scenario and Tasks involved

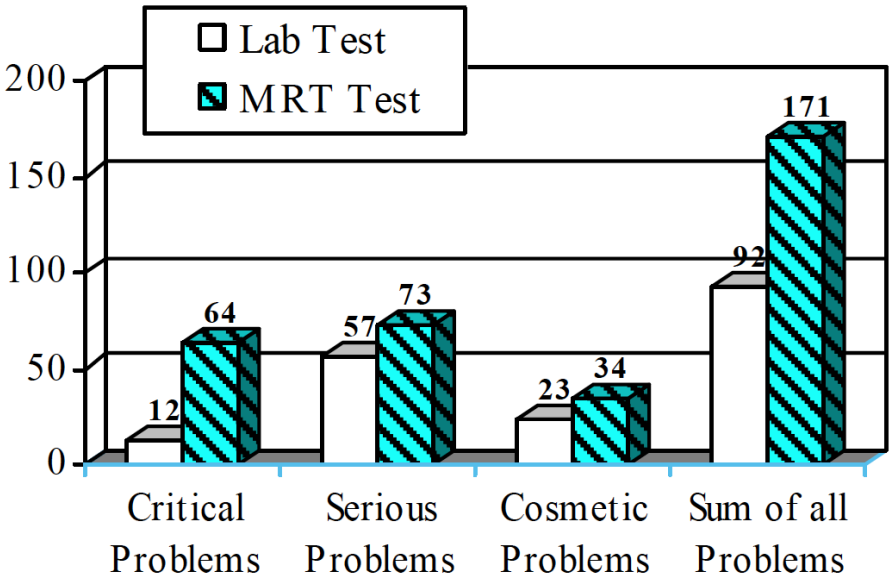
Task	Scenario of action	Task Description
1	You need to inform your friend about your personal particulars as he needs to fill up a form for you. You decide to call out.	1. Dial out to contact Gerald from mobile phone contacts list. 2. Start a conversation upon pick up as you normally would. 3. Verbally inform the contact your full name, NRIC, address and date of birth
2	You receive a call from a friend on your mobile phone. You answer the phone call.	1. Answer phone call as you normally would. 2. Start a conversation with the friend.
3	You need to inform your friend about your personal particulars information as he needs to fill up a form for you. You decided to SMS	1. Compose a SMS including the following information: your full name, NRIC, address and date of birth. 2. Send SMS to Gerald from mobile phone contact list. 3. Reply again to Gerald if necessary, i.e. if Gerald replied your message.

Image source: Duh, Tan, Chen: Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. MobileHCI 2006.

Comparison of Lab and Field Tests

universitätbonn

Problems found:



User behavior:

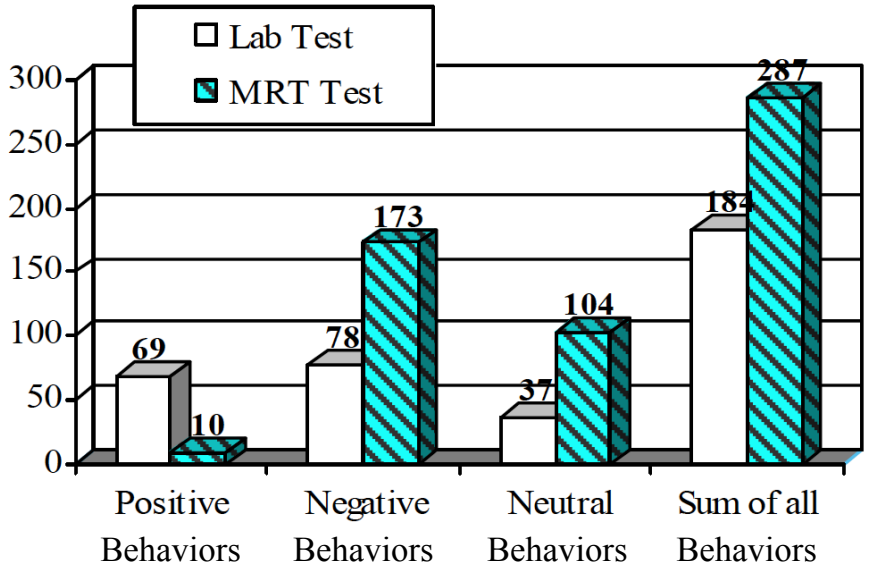


Image sources: Duh, Tan, Chen: Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. MobileHCI 2006.

Evaluating Attentional Resources in Mobile HCI

- Evaluating the competition for cognitive resources when mobile
- Field study in urban environment
  - Performance of mobile Web tasks
  - Movement through urban situations
- Attention during loading a page
  - Duration of continuous attention
    - Lab: 16.2s → field: 4s
  - Number of attention switches
    - Lab: 1 → field: 8
  - Switching-back durations
    - Railway station: 7-8s, quiet street: 4-6s

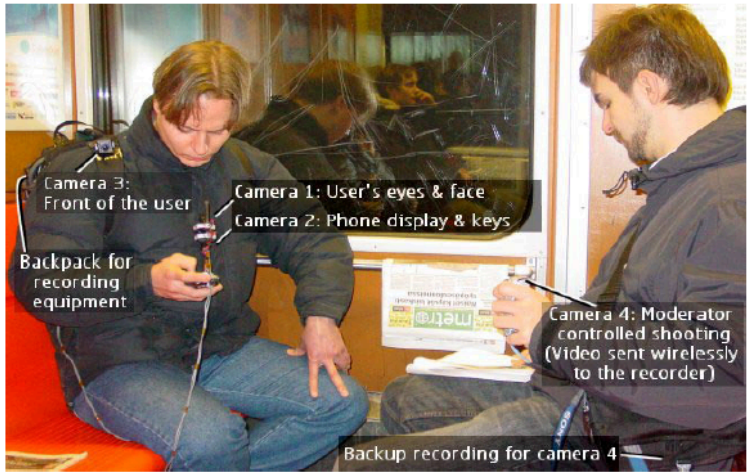
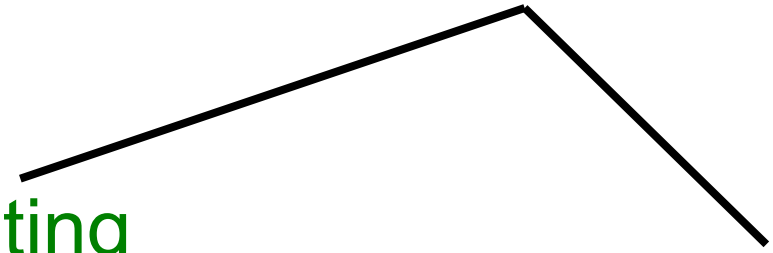


Figure 2. Configuration of recording equipment.

Figure 3. Output video data integrated on-the-fly.

Oulasvirta, Tamminen, Roto, Kuorelahti. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. CHI '05.

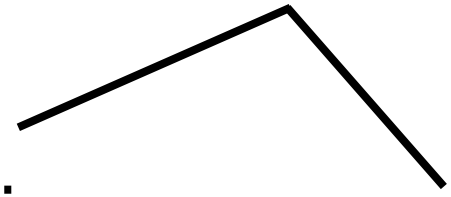
Evaluation Techniques



Evaluating Without Users

E1 Literature Review
E2 Cognitive Walkthrough
E3 Heuristic Evaluation
E4 Model-Based Evaluation

Evaluating With Users



Qualitative

E5 Conceptual Model Extraction
E6 Silent Observation
E7 Think Aloud
E8 Constructive Interaction
E9 Retrospective Testing

+ Interviews, questionnaires,...

Quantitative

E10 Controlled Experiments

E1: Literature Review

- Many research results about user interface design have been published
- Idea: Search literature for evidence for (or against) aspects of your design

\+ Saves effort, time and money by avoiding own experiments

– Results only carry over reliably if context (users, assumptions) is very similar

E2: Cognitive Walkthrough

- Analytical method for early design or existing systems
  - Without users
- Expert evaluator = designer or cognitive psychologist
- Goal: Judge learnability and ease of use
  - Does system help user to get from goals to intentions and actions?
- Step through each action and ask
  - Is the effect of the action the same as the user's goal at that point?
  - Will users see that the action is available?
  - Once users find the action, will they know it is the right one?
  - After the action is taken, will users understand the feedback?

E2: Cognitive Walkthrough

- What you need
  - Interface description (prototype of the system)
  - Task description
    - Example: Program the DVR to time-record a program starting at 18:00
      and finishing at 19:15
      on BBC 1 on June 2, 2018
  - List of interface actions to complete the task
  - User profile
- Doing the actual walkthrough
  - Analyze process of performing the actions using above questions
- Written questions capture psychological knowledge and guide the tester

E3: Heuristic Evaluation

- Choose usability heuristics
  - (general usability principles, e.g., Nielsen's 10 Usability Principles)
- Step through tasks and check whether guidelines are followed
- Severity rating for each problem (Nielsen)
  - 0 = I don't agree this is a problem at all
  - 1 = cosmetic problem
  - 2 = minor usability problem, low priority to fix
  - 3 = major usability problem, high priority to fix
  - 4 = usability catastrophe, imperative to fix before release

\+ Quick and cheap

– Subjective (have several independent evaluators)

See also: www.useit.com/papers/heuristic

10 Usability Principles (Jakob Nielsen)

1. Keep the interface simple!
2. Speak the user's language!
3. Minimize the user's memory load!
4. Be consistent and predictable!
5. Provide feedback!
6. Design clear exits and closed dialogs!
7. Offer shortcuts for experts!
8. Help to recover from errors, offer Undo!
9. Prevent errors!
10. Include help and documentation!

10 Rules for a good Crypto API? Smith & Green @ USENIX Hotsec'15

universitätbonn

1. Easy to learn, even without crypto background
2. Easy to use, even without documentation
3. Hard to misuse. Incorrect use should lead to visible errors
4. Hard to circumvent errors – except during testing/development
5. Easy to read and maintain code that uses it
6. Sufficiently powerful to satisfy (non-security) requirements
7. ~~Easy to extend~~ Hard to change/override core functionality
8. Appropriate to audience – this means people with no crypto experience
9. Assist with/handle end-user interaction

conduct developer studies

10. However, where possible integrate into standard APIs so normal developers never have to interact with crypto APIs in the first place

Categories of Design Rules

universitätbonn

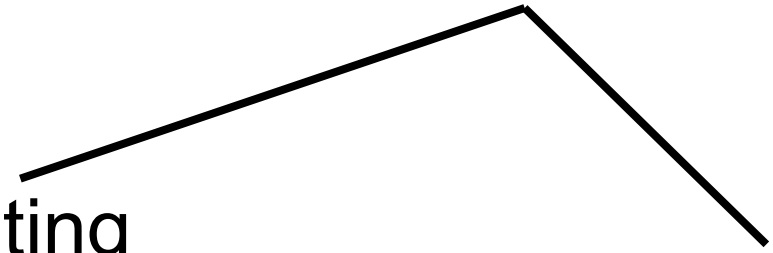
- Principles, Heuristics
  - Small set of general rules (low authority, high generality)
  - Abstract rules, based on psychological knowledge
  - Largely independent of technology
- Guidelines
  - Large set of detailed rules (medium authority, low generality)
  - Often developed for a specific platform
  - More concrete, more technology-oriented
- Standards
  - Agreed upon by a large community (high authority, medium generality)
  - Carefully developed by a standards committee (consensus-based)

Standards

Examples:

- ISO 9241: “Ergonomics of Human System Interaction”, 17 parts
  - 7 parts concerning hardware issues, 8 parts concerning software issues
- ISO 14915: “Software ergonomics for multimedia user interfaces”, 3 parts
  - “Multimedia navigation and control”, “Media selection and combination”

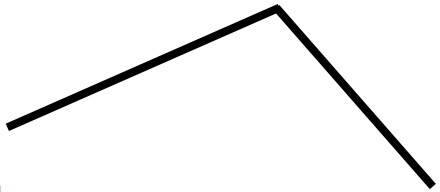
Evaluation Techniques



Evaluating Without Users

E1 Literature Review
E2 Cognitive Walkthrough
E3 Heuristic Evaluation
E4 Model-Based Evaluation

Evaluating With Users



Qualitative

E5 Conceptual Model Extraction
E6 Silent Observation
E7 Think Aloud
E8 Constructive Interaction
E9 Retrospective Testing

Quantitative

E10 Controlled Experiments

+ Interviews, questionnaires,...

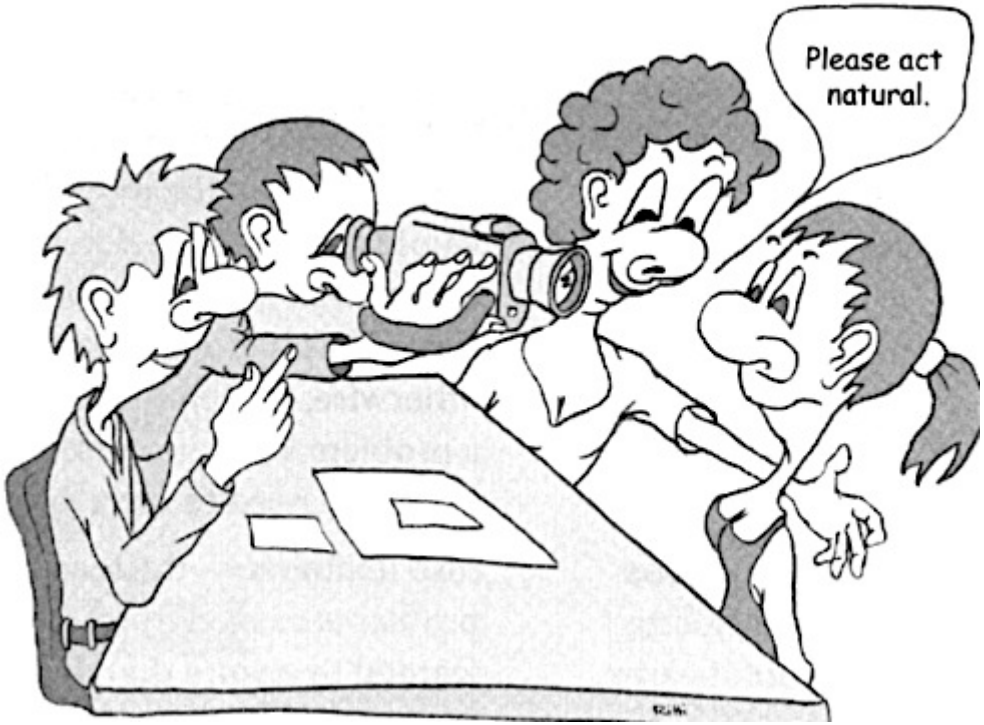
Four Key Issues

1. Setting goals
   - Decide how to analyze data once collected
2. Relationship with participants
   - Clear and professional
   - Protect privacy
   - Informed consent-form
     - Signed agreement between evaluator and participant
3. Triangulation
   - Use more than one approach
   - Use different perspectives to understand a problem or situation
4. Iterate
   - If questions reveal that goal was not sufficiently refined: refine goal, repeat

Dealing with Test Users

- Tests are uncomfortable for the tester
  - Pressure to perform, mistakes, competitive thinking

- So treat testers with respect at all times!
  - Before, during, and after the test



Recording Observations

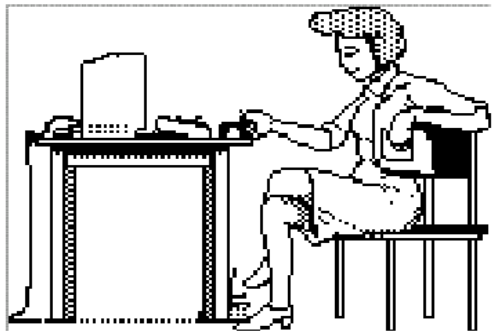
- Paper and pencil
  - Evaluator notes events, interpretations, other observations
  - Cheap but hard with many details (writing is slow)
  - Forms can help
- Audio recording
  - Good for speech with Think Aloud and Constructive Interaction
  - But hard to connect to interface state
- Video
  - Camera on user +
  - Screen-capture
  - Best capture, but may be too intrusive initially
  - Time consuming to process during evaluation
- Logging
  - Log input events of the user, synchronize with audio & video

Before you begin

universitätbonn

- Test
- Test
- And test again

Silent Observation



Source: Saul Greenberg

- Designer watches user in lab or in natural environment while working on one of the tasks
- No communication during observation

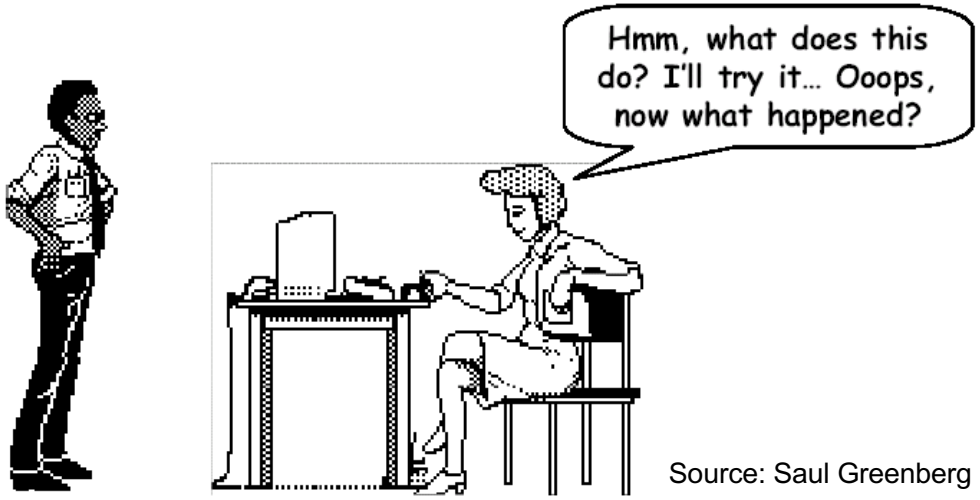
\+ Helps discover big problems

\+ Prevents “overly helpful” assistant problem

– No understanding of decision process (that may be wrong) or user’s mental model, opinions, or feelings

– Can end in almost zero result, if the participant gets things wrong in the beginning

Think Aloud



Source: Saul Greenberg

- As Silent Observation, but user is asked to say aloud
  - What he thinks is happening (state)
  - What he is trying to achieve (goals)
  - Why he is doing something specific (actions)
- Most common method in industry

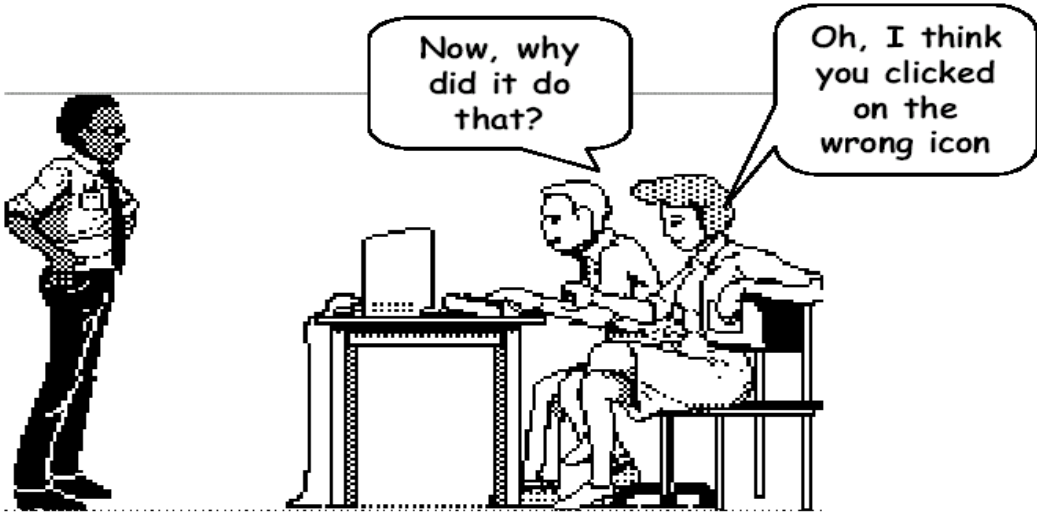
\+ Good to get some insight into user's thinking, but:

– Talking is hard while focusing on a task

– Feels weird for most users to talk aloud

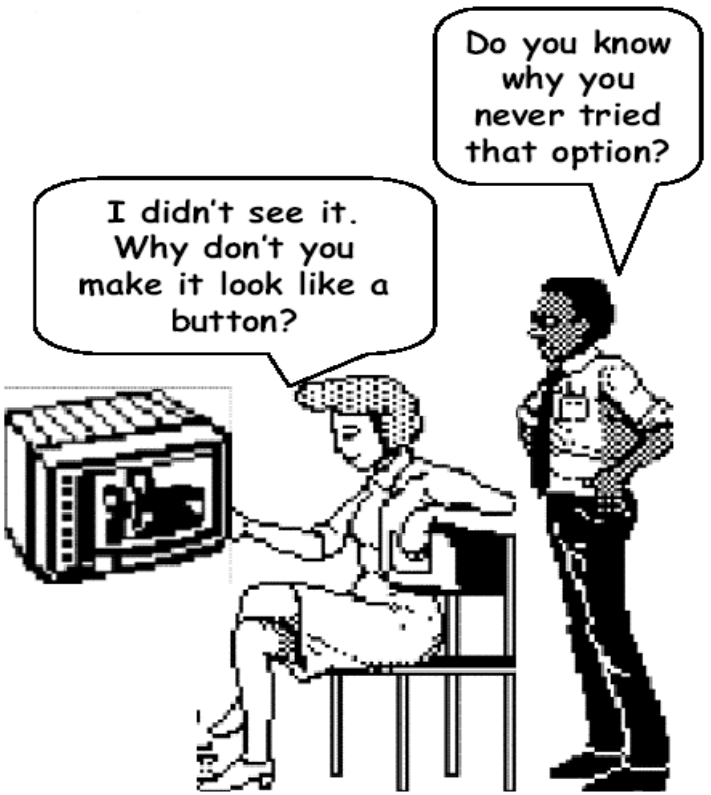
– Conscious talking can change behavior

E8: Constructive Interaction



- Two people work on a task together
  - Normal conversation is observed (and recorded)
  - \+ More comfortable than Think Aloud
- Variant of this: Different partners
  - Semi-expert as “trainer”, newbie as “student”
  - Student uses UI and asks, trainer answers
  - \+ Gives insight into mental models of beginner and advanced users at the same time!

E9: Retrospective Testing



- Additional activity after an observation
- Subject and evaluator look at video recordings together, user comments his actions retrospectively
- Good starting point for subsequent interview, looking at video avoids wrong memories
- Often results in concrete suggestions for improvement

INTERVIEWS

Pros and Cons of Interviews

+ Ability to go deep
  - by asking that explore a wide range of concern
  - and giving interviewees the freedom to provide detailed responses
  - data can be gathered that would otherwise be very hard to capture
– This flexibility comes at a price:
  - potentially unbounded discussion must be managed
  - interviews are harder to conduct than surveys
  - time consuming
– Analysis is a major challenge
  - transforming and merging raw notes into usable data is challenging and time consuming
– Interview is separate from task
  - problem of recall
  - self-reported data

The Interview

- Tell the interviewee that he/she can decline to answer any question or opt-out entirely at any time
- State what the study is about (+/- subterfuge)
- Start with relatively easy questions
  - this builds trust and confidence
- Intersperse hard questions with easy questions
  - to defuse any tension
- Critical questions should be done near the end
- End the interview with easy questions
  - this creates the feeling of accomplishment
- Debrief the interviewee and/or offer a follow-up information exchange
- Thank the interviewee for their time and participation

Debriefing

- Turn of recording devices for debriefing
  - participants might share comments they were not willing to say during the interview
  - care is needed in dealing with this data!

Try it out

- Mini interview
- Please form groups of two
  - one interviewer
  - one interviewee
- Task:
  - conduct an interview to learn about the technical procedure how to construct a peanut butter and honey sandwich
  - assuming that you have new jars of peanut butter and honey, a fresh loaf of bread and a standard kitchen
  - find as many technical details as possible!
  - use your smart phones to record the interview (optional)

Discussion

- Get plate
- Get cutlery
- Open bread packaging
  - were you asked if the bread came in packing?
  - did you have to decide on the fly?
- Cut bread with bread knife
- Open jars
- Remove foil
  - this is often forgotten since it is implicit knowledge
- Spread peanut butter and honey
  - preferences for order?
  - spoon for the honey?
  - same knife for both?
  - single slice or double decker?

Types of Interview

- Unstructured / Open ended / Exploratory
  - Some specific questions that are planed
  - But based on interviewees responses interviewers can
    - re-order questions
    - invent new lines of inquiry
- Structured
  - rigid script with the questions
  - questions can be skipped based on pre-defined rules
  - similar to survey
    - but it might be easier to answer a question verbally than to write it down
  - easier to analyse
- Semi-Structured
  - mix of the above

Types of Questions

- Closed questions
  - yes-no
  - multiple choice
  - likert scale
  - Do you like the design of this warning message?
- Open questions
  - Questions asking for responses, opinions, feedback without external constraints
  - What did you think about the design of this warning message?
- Tasks (hidden questions)
  - Please complete this sentence:
  - “The most frustrating problem with PGP is…”
- Conceptual Mapping
  - “Please draw a diagram showing how web security works”

Keep it simple

- Avoid compound questions
  - – What were the strengths and weakness of the menu and the toolbar?
  - + What did you think about the menu layout?
  - + What did you think of the toolbar?
  - + Which did you prefer?

- Paraphrase complex or unclear answers
  - “Did I understand you correctly, you think that…”
  - The closed yes/no questions allow you to be sure that you extracted the right information

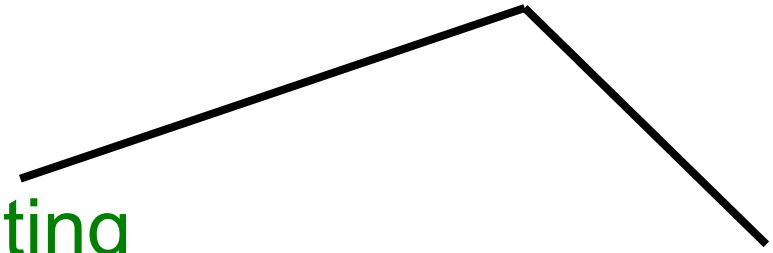
Try it out

- Mini interview
- Please form groups of three (7 min., then change roles)
  - one interviewer
  - one interviewee
  - one referee
- Task:
  - Find out the details of a problem of choice of the interviewee
  - Examples
    - Problems of using PGP
    - Trust in Crypto-currencies
  - Only use open questions!
  - If closed question is asked interviewee answers only with yes or no and says nothing else
  - Referee buzzes when these rules are broken
  - Use paraphrases to check if you understood correctly

Avoid Bias

- Questions should be as unjudgmental and unbiased as possible
- Watch out for phrasing that could encourage participants to give answer they think you want to hear
  - Particularly if you are asking about something you built
- Examples:
  - – Why do you like this design?
  - – Don’t you think this is difficult to use?
  - – Did you like…?
  - + What did you think of…?

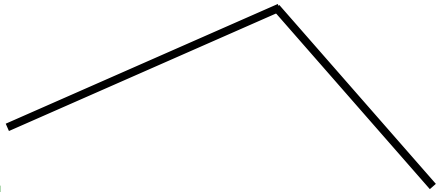
Evaluation Techniques



Evaluating Without Users

E1 Literature Review
E2 Cognitive Walkthrough
E3 Heuristic Evaluation
E4 Model-Based Evaluation

Evaluating With Users



Qualitative

E5 Conceptual Model Extraction
E6 Silent Observation
E7 Think Aloud
E8 Constructive Interaction
E9 Retrospective Testing

+ Interviews, questionnaires,...

Quantitative

E10 Controlled Experiments

E10: Controlled Experiments

universitätbonn

- Quantitative, empirical method
- Steps
  - Formulate hypothesis
  - Design experiment, pick variable(s) and fixed parameters
  - Choose subjects
  - Run experiment
  - Interpret results to accept or reject hypothesis

Briefly: Statistics Nomenclature

- Statistical Unit
  - Objects from which measurements are collected (e.g. a participant or a country)
- Population
  - The set of all statistical units relevant to a particular investigation
- Sample
  - The subset of the population that was actually analysed
- Attribute/Parameter/Variable
  - A property of the statistical unit that we are interested in
- Value
  - The actual data for a variable measured for one statistical unit

Two Main Types of Variables

- Independent (IV) / Predictor / Factor / Input
  - what we base our explanation on
  - characterize statistical units
  - examples: age, expertise

- Dependent (DV) / Outcome / Target / Output
  - what we are trying to explain
  - examples: usability, time taken

Hypothesis

- A claim that predicts outcome of a DV based on an IV
- Approach: Reject null hypothesis (inverse, i.e., “no influence”)
  - Null hypothesis is a term from statistical testing

- Consider this hypothesis:

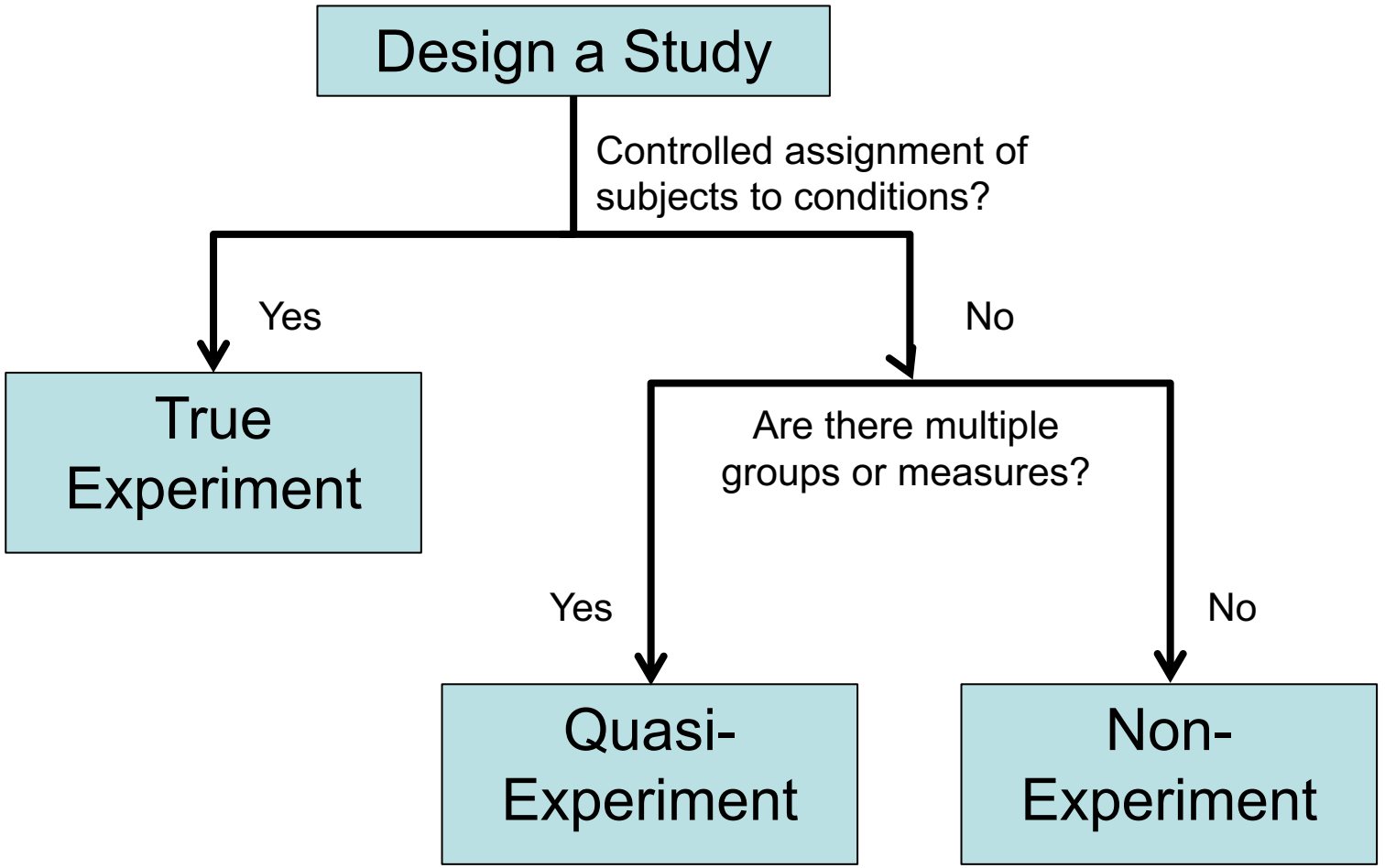
  “There is no difference between the target selection speed when using a mouse, a joystick, or a trackball to select icons of different size (small, medium, large)”.

  - What are the DV and the IVs? How many conditions in the experiment?

Conditions

- Condition
  - The procedure that is varied in order to estimate a variable's effect by comparison with a control condition
- Number of conditions
  - product of the number of values in each IV
  - in our example: 3 x 3 = 9
- How to control the IVs or condition assignment?
  - straightforward in previous example: get joystick, mouse or trackball
  - challenging in other cases:
    - testing influence of demographic properties (cognitive abilities, left-handed vs. right-handed, …)
    - testing against best-case scenarios that do not exist (Wizard of Oz).

Types of Research Designs



Types of Research Designs

- True Experiment
  - the experimenter controls assignment of experimental units (e.g., participants, rats) to experimental conditions
  - control allows to draw causal conclusions
  - example:
    - randomly give half of the participants the real drug and the other half a placebo
    - let half of the participants randomly use one product or the other and measure usability.

Types of Research Designs

- Quasi-Experimental
  - if assignment cannot be controlled, another assignment criterion is used
  - examples:
    - some statistical units occurred before a certain *“external”* event and some after (movie revenue before and after Megaupload was closed down)
    - compare pupils according to grade averages above and below a threshold or from one type of school to another
    - compare people who have a security background with those who don’t
  - Remember: more than one condition is needed to be a quasi-experiment

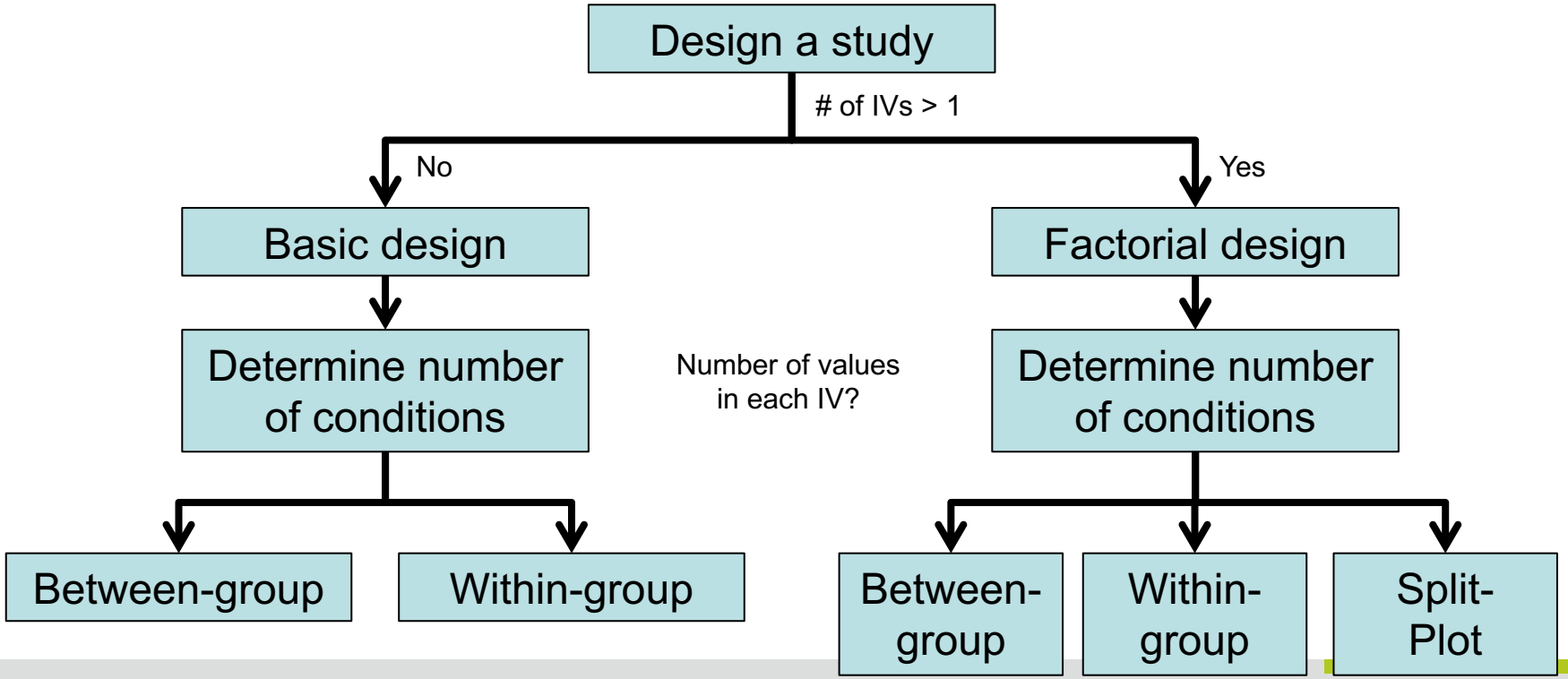
Types of Research Designs II

- Non-Experimental
  - describe phenomenon “as is”
  - do not manipulate variables
    - therefore, cannot deduct a cause!
  - e.g. surveys, ethnography(, interviews), …

- These three forms can be mixed, especially when conducted online.

Basic Design Structure

- Goal: draw a big picture of how to run the experiment
  - estimate a timeline (and a budget)
- Two essential questions:
  - How many IVs do we want to investigate in this experiment?
  - How many different values does each IV have?



Single-IV Experiments

- Consider the following hypotheses. How many conditions in each hypothesis?

  - H1: “There is no difference in typing speed when using a QWERTY keyboard, a DVORAK keyboard, or an alphabetically ordered keyboard.”
  - H2: “There is no difference in the time required to locate an item in an online store between novice users and experienced users.”
  - H3: “There is no difference in the perceived trust towards an online agent among novice and experienced customers who are from the United States, Russia, China, and Nigeria.”

Participant Exposure

- How many conditions do we expose each participant to?
  - Between-group / between-subject design
    - The effect of each condition is measured between groups/subjects.
      - i.e. each participant is exposed to one condition only.
    - If the task is to type 500 words using a selected keyboard, each participant types 500 words.
  - Within-group / within-subjects design
    - The effect of each condition is measured within the group/each subject.
      - i.e. one group of participants is exposed to all conditions
    - In this case, each participants types 1500 words.
- This decision implies the use of different statistical analyses.

Active Learning

- Think – Pair – Share
- What are the advantages/disadvantages/differences between
  - Within subjects
  - Between subjects
- ?

Between-group Design

- Cleaner design
  - no learning from previous exposures
  - less time spent in the experiment
    - ➔ less influence of fatigue and frustration

- Compare two distinct groups of participants
  - there is no baseline for every individual
    - individual differences cause noise
    - need to make sure groups are very similar
  - number of participants in each group needs to be high
    - sample size = no. of conditions x no. of participants per condition (as dictated e.g. by power analysis).
    - example: 4 conditions x 16 participants/condition: N=64

Within-group Design

- Exactly the opposite of between-groups design:
  - smaller sample size needed: N=16 (!)
    - therefore usually a lot cheaper to conduct
  - more meaningful measurements due to individual comparisons

- Disadvantages
  - repeated exposure can cause learning and fatigue
    - learning favours subsequent exposures
    - fatigue favours initial exposures
    - can add to substantial overall bias

When to use Between-group Design

- Difficult decision, to be made on case-by-case basis
- Hybrid setups possible (later slides)
- Between-groups should be used for:
  - simple tasks with limited individual differences
    - limited cognitive processes
    - e.g. basic motor skills when selecting a screen target
      - as opposed to reading, comprehension, information retrieval and problem solving
  - tasks that would be greatly influenced by learning effects
    - first-contact required, e.g. when testing website design
  - problems that cannot be investigated using within-groups design
    - consider H2 and H3 from before

Using Between-group Design

- Randomly assign participants to conditions
  - randomly does not mean haphazardly!

- Counterbalance confounding factors
  - gender
  - age
  - computing experience
  - Internet experience
  - …
  - i.e. all (“relevant”) demographic properties except those that are IVs

- Make sure groups are as similar as possible w.r.t. your hypothesis.

When to use Within-group Design

- Within-groups design isolates individual differences more effectively
- Within-groups should be used for:
  - tasks with large individual differences
    - i.e. reading, comprehension, information retrieval and problem solving
  - tasks that are less susceptible to learning effects
    - involving complicated or learnt skills and knowledge
    - for example: investigating the impact of font type on reading speed
  - very small target participant pools
    - for example, when looking for a particular participant property (disabilities, illnesses, or a combination of demographical properties)

Using Within-groups Design

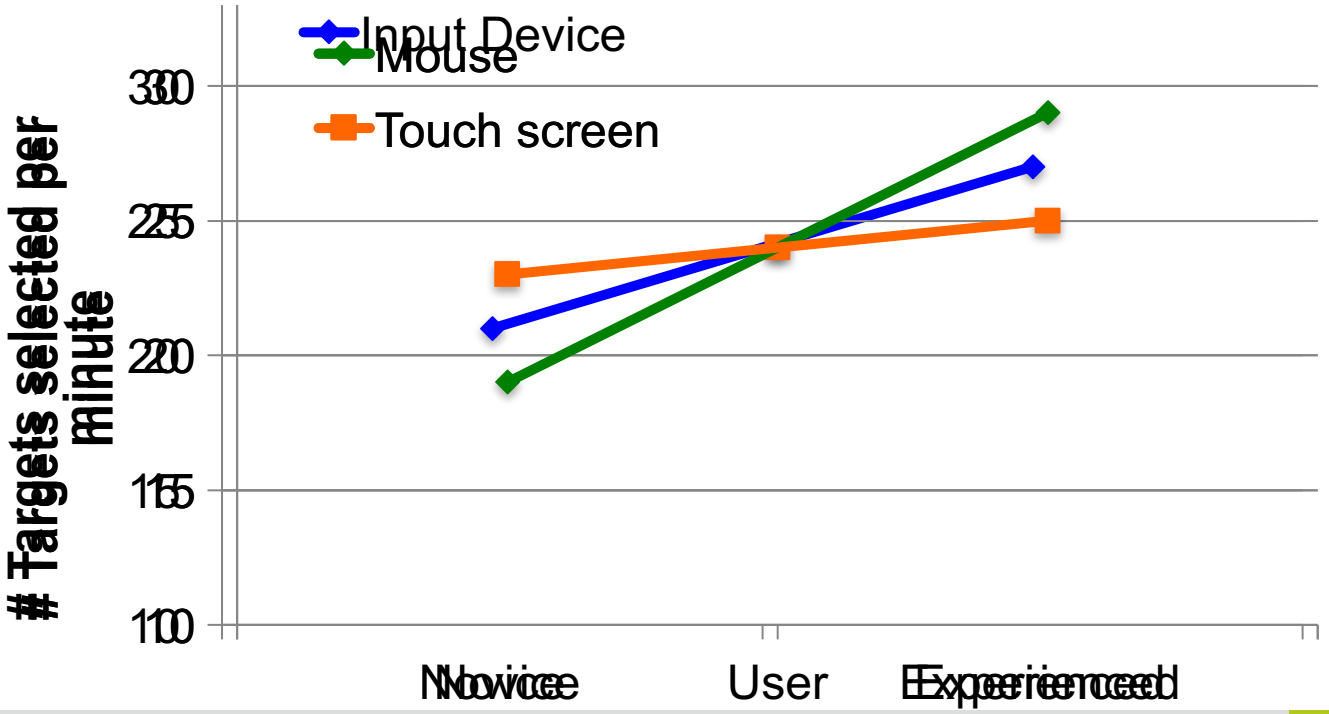
- Need to control for negative impact of learning, fatigue and other within-groups problems
  - randomise task order to control for learning
    - for example, learning effects in one participant using the DVORAK keyboard last are offset by another participant using it first
  - provide a training session
    - if participants can familiarise with all conditions before the actual experiment, learning has less influence
    - commonly used in combination with task randomisation
  - limit the total time spent in the experiment
    - generally between 60 and 90 minutes or less
    - never more than 120 minutes
    - plan (force) breaks when necessary

Factorial Designs

- Used to investigate more than one IV
  - number of conditions is the product of the number of values in each IV
  - example: in addition to three **keyboard types** (QUERTY, DVORAK, and Alphabetic), we also want to investigate the effects of **different tasks** (composition vs. transcription) on typing speed.
  - 3 x 2 = 6 conditions
    - across the **task dimension**, we can examine the impact of keyboards
    - across the **keyboard dimension**, we can examine the impact of task type
    - across **both IV dimensions**, we can examine interaction effects

Interaction Effects

- "The effect of one IV on the DV, depending on the particular value of another IV".
  - one IV alone may not cause significant effects
  - interaction effects can provide additional insights
- Example (IVs: input device and expertise, DV: selection speed)

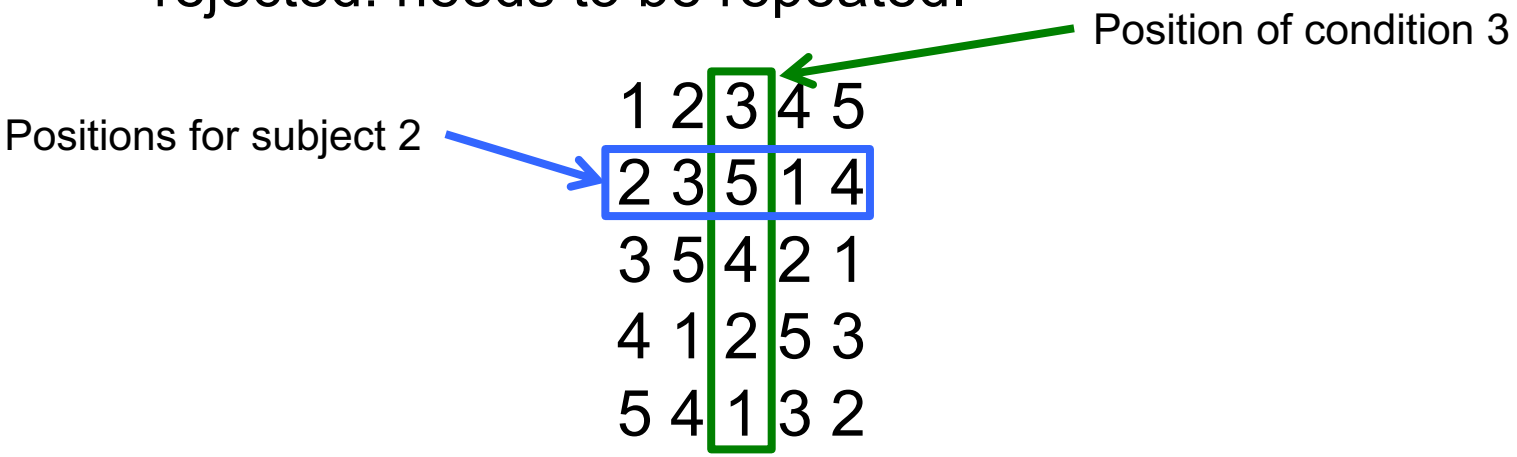


Split-Plot Design

- Investigate some variables within-groups and others between-groups
  - example: testing the influence of age and GPS assistance to driving efficiency.
    - age cannot be tested within-groups
    - but each driver can drive with and without GPS assistance
  - advantage: smaller sample but still baseline for each participant in some IVs

Latin Square Design

- Variation of Split-Plot: Latin Squares
  - Latin squares make sure that each condition is assigned to each position the same number of times in a within-subjects study.
  - simulates a between-subjects design on each position
  - is not a truly random assignment!
  - can be difficult to administer
    - Latin square property is broken when one record is rejected: needs to be repeated.



Reliability of Experimental Results

universitätbonn

- Reliable Experiments can be replicated by others teams, in other locations and another time
  - yielding results that are consistent, dependable, and stable

- Big challenge for HCI compared to “hard sciences”
  - measurements of human behaviour and social interaction subject to higher fluctuation and individual differences
  - less replicable
  - especially when using smaller samples

- Fluctuations in experimental results: Errors
  - okay: random errors
  - bad: systematic errors!

Random Errors

- If observing someone typing documents in five sessions, we may obtain a typing speed of 50 words per minute
  - in each session, speeds will vary:
    - 46 words/min, 52 w/m, 47 w/m, 51 w/m, 53 w/m
    - on average: about 50 words/min

- Observed values = actual value + random error
  - random error = noise
  - occurring by chance
  - push actual value up and down
  - the larger the sample, the more the noise equalises
  - therefore, can be ignored

Systematic Errors

- Non-random error
  - also called bias
  - systematically influencing measurement values
    - pushing values into the same direction
  - no offsetting in larger samples
  - typing example:
    - imagine the experimenter is looking over the secretary's shoulder and then gets these results:
      - 47,44,45,42,46 ➔ mean=44.8
    - still fluctuating, but underperforming

Systematic Errors

Instability	Biasedness: Much	Some	None
Much			
Some			
None			

- Instability of sampling vs. bias (Rosnow & Rosenthal, 2008)
- Bias is the true enemy of experimental research
  - try to eliminate and control bias
  - isolate impact when inevitable
  - five major sources

Measurement Instrument Bias

- Instruments used to measure values can introduce systematic error
  - e.g. ill-calibrated sensor, slow stopwatch, misused questionnaire

- Can be avoided by carefully selecting and examining instruments before the experiment

Experimental Procedure Bias

- Inappropriate or unclear experimental procedures may introduce biases
  - learning, fatigue
  - wording of instructions
    - need to be consistent across conditions and instructors!
    - example: participants instructed to work "*as quickly as possible*" were slower than those instructed to "*take your time, there is no rush*".
  - ambiguous instructions can cause unwanted variations:
    - e.g., some participants may hold a smartphone sideways and be quicker than those who hold it regularly
    - can be avoided by clarifying instructions
- Pre-testing is essential!

Experimental Procedure Bias

- To reduce bias through experimental procedures
  - randomise the order of conditions, tasks, and task scenarios in within-group and split-plot designs
  - prepare a written document with detailed, tested and revised instructions for participants
  - run pilot studies/pre-tests with actual participants before collecting real data and revise your procedure accordingly
  - prepare a written document with detailed, tested and revised instructions for experimenters
    - even if you are the only experimenter!

Participant and Recruiting Bias

- Consider participant demographics and possible biases
  - e.g., recruiting IT students only will bias your results!

- Recruit carefully: make sure your participant pool mirrors the target user population (as best as possible)
- Create an environment and task procedure that causes the least stress to the users
- Reassure participants that you are testing a product/an interface/a method and not them.
  - has been shown to make people more calm and relaxed during experiments
- Be prepared to reschedule exhausted, stressed, tired or very nervous participants.

Experimenter Behaviour Bias

- Major source of bias
  - may intentionally or unintentionally influence experiment results
    - worse if multiple experimenters are used: different influences are hard to discover and quantify
  - spoken language, body language, and facial expressions cause bias!
    - “I think you will like this… I designed it myself!”
    - a frustrated looking experimenter says: “Damn! That’s so slow!”
    - a waiting experimenter leans forward and taps his fingers on the mouse while waiting for an application to load.
    - the experimenter arrives late and after the participant. He then takes 10 minutes to set everything up…

Experimenter Behaviour Bias

- Controlling for experimenter behaviour bias:
  - train experimenters or practice yourself
    - stay neutral, calm and patient during experiments no matter what happens
  - experimenters should make sure to arrive 15 minutes ahead of scheduled sessions to prepare everything.
  - supervise a session using two experimenters when possible.
    - lead experimenter interacts with participants, assistant observes and takes care of technical problems.
  - prepare written documents with detailed procedures that all experimenters have to follow strictly.
  - when appropriate, record instructions for the participants before the experiment and play the recording during the experiment.

Environmental Factors Bias

- Make sure the environment is agreeable
  - quiet room, appropriate lighting, comfortable chairs and tables, clean, tidy, no distractions.
  - if possible, the participant should be seated alone and the experimenter observes from another room.
  - in a field study, check the study location before the scheduled time.

- A final note on bias: we can try to avoid bias, but there always will be some bias.
  - We should be careful when interpreting and reporting the findings.
  - We should be ready to argue why a certain bias is acceptable.
    - for example: using only students

Choosing a Method

- Between-groups
  - Each subject only does one variant of the experiment
  - There are at least 2 variants (manipulated form & control, to isolate effect of manipulation)
  - \+ No learning effect across variants
  - − But requires more users
- Within-groups
  - Each subject does all variants of the experiment
  - \+ Less users required, individual differences canceled out
  - − But often learning effect across variants problem

Analyzing Results

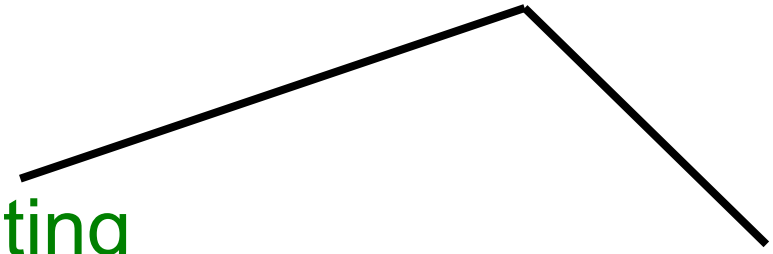
- Statistical analysis
  - Often assumptions about underlying distribution
  - t-test: Compare two groups, normal distribution
  - Analysis of variance (ANOVA): Compare two or more groups, normal distribution
  - Regression analysis: How well does result fit to a model?
  - Wilcoxon- or Mann/Whitney test, X^2 test
- Choice depends on
  - Number, continuity, and assumed distribution of dependent variables
  - Desired form of the result (yes/no, size of difference, confidence of estimate)

Other Evaluation Methods

universitätbonn

- Before and during the design, with users
  - Personal interviews
  - Questionnaires
- After completing a project
  - Email bug report forms
  - Hotlines
  - Retrospective interviews and questionnaires
  - Field observations (observe running system in real use)

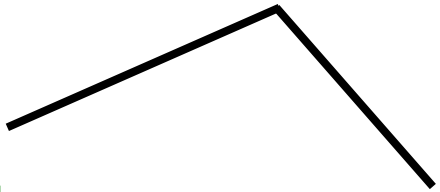
Evaluation Techniques



Evaluating Without Users

E1 Literature Review
E2 Cognitive Walkthrough
E3 Heuristic Evaluation
E4 Model-Based Evaluation

Evaluating With Users



Qualitative

E5 Conceptual Model Extraction
E6 Silent Observation
E7 Think Aloud
E8 Constructive Interaction
E9 Retrospective Testing

+ Interviews, questionnaires,...

Quantitative

E10 Controlled Experiments

Summary

universitätbonn

- Evaluate to ensure system matches users' needs
- Evaluation should happen throughout the design process
  - By experts (analytically)
  - By users (experimentally)
- A plethora of methods to evaluate designs
  - Decide when to apply which
- Treat testers with respect at all times!