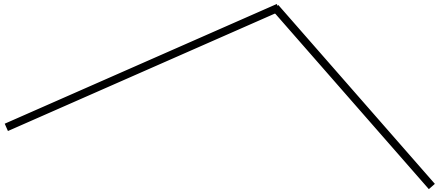
Quantitative Studies

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

Quantitative, empirical method

Steps

- Formulate hypothesis
- Design experiment,
  - pick variable(s)
  - fix parameters
  - plan data analysis
- Recruit subjects
- Run experiment
- Interpret results to accept or reject hypothesis

Hypothesis

Hypothesis

- A claim that predicts outcome of experiment
  - Example: Warning A leads to lower click-through rates (CTR) than Warning B
- Hypothesis claims that changing independent variables influences dependent variables
  - Example: Showing Warning A (independent variable) influences CTR (dependent variable)
- Experimental goal: Confirm hypothesis
- Approach: Reject null hypothesis (inverse, i.e., “no influence”)
  - Null hypothesis is a term from statistical testing: The samples are drawn from the same statistical distribution

Variables

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

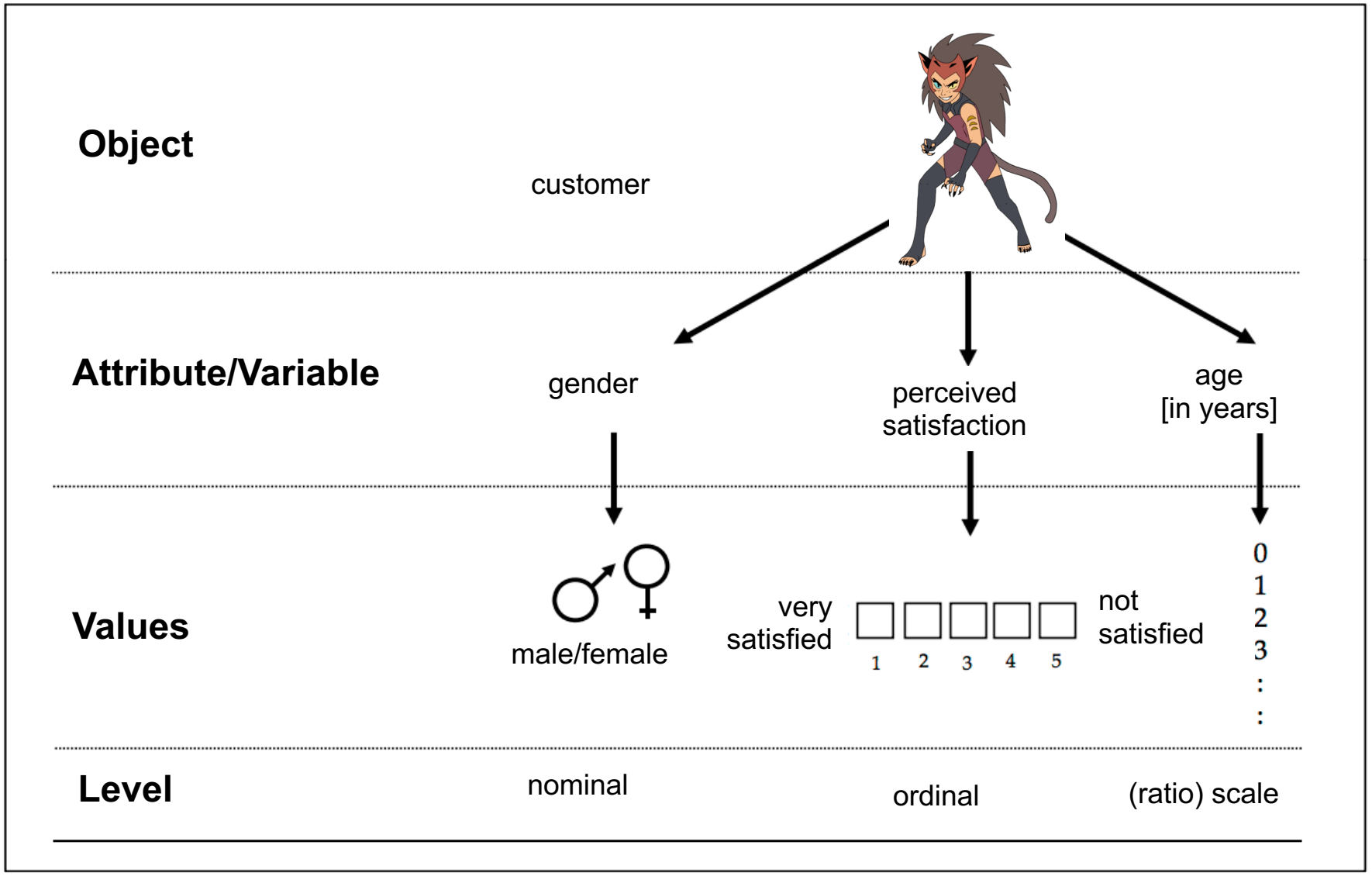
Example Independent Variables

- User Interface
  - Warning A, Warning B, Warning C
  - old HTTPS indicator, new HTTPS indicator
  - PGP 9.0 plugin, PGP 10.0 plugin
  - Documentation A, Documentation B
- Programming Interfaces
  - Spring API, JSF, Bouncycastle
- Software
  - AFL, libFuzzer
- Processes
  - SCRUM, waterfall model

Example Dependent Variables

- Time taken
- Mistakes made
- Tasks solved
- Satisfaction
- etc. see on measuring usability

Variables composition



Levels of Measurement

- Nominal
  - discrete, but unrelated categories, equal or unequal
  - example: day/night, type of cake
- Binary/dichotomous
  - special case of nominal/ordinal, frequently used
- Ordinal
  - discrete categories that can be ordered
  - example: 5/7 point scales, small/medium/large pizza
- Interval scale: discrete or continuous measurements where distance makes sense
  - example: Temperature in Celsius, often ratings
- Ratio scale: Interval + true zero (ratios make sense)
  - example: Time, Age, Money, Distance, Temperature in Kelvin, Lines of code

Example Malware Study

3 Independent Variables

- Decompiler
  - Hex-Rays
  - DREAM
  - DREAM++
- Participant
  - Student
  - Expert
- Task
  - Medium
  - Difficult

Dependent Variables

- Tasks completed
  - ordinal
- Time taken
  - ratio
- Satisfaction
  - ordinal
- Trust
  - ordinal

Conditions

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 2 x 2 = 12
- Conditions:
  - Hex-rays, Experts, Medium
  - Hex-rays, Experts, Hard
  - Hex-rays, Students, Medium
  - Hex-rays, Students, Hard
  - DREAM, Expert, Medium
  - …

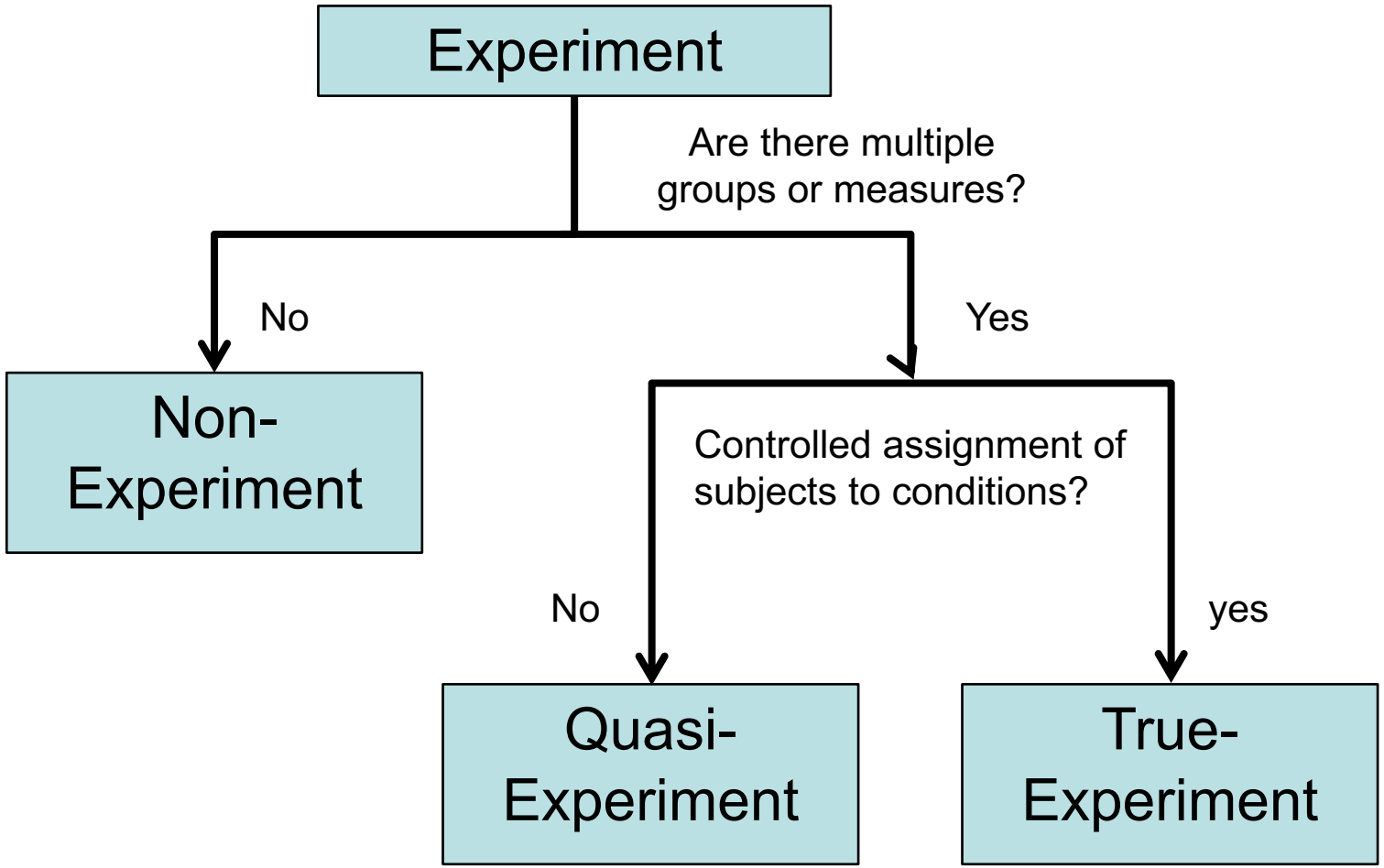
Conditions

How to control the IVs or condition assignment?

- straightforward in some cases
  - randomly assign interface to participant
- challenging in other cases:
  - demographic properties (gender, income, education, etc.)
  - complex skills (Linux vs Windows user, expert vs beginner)
  - events (pre-snowden vs post-snowden)

Experiments

Types of Research Designs



Types of Research Designs

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

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 (privacy attitude before and after Snowden revelations)
    - 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

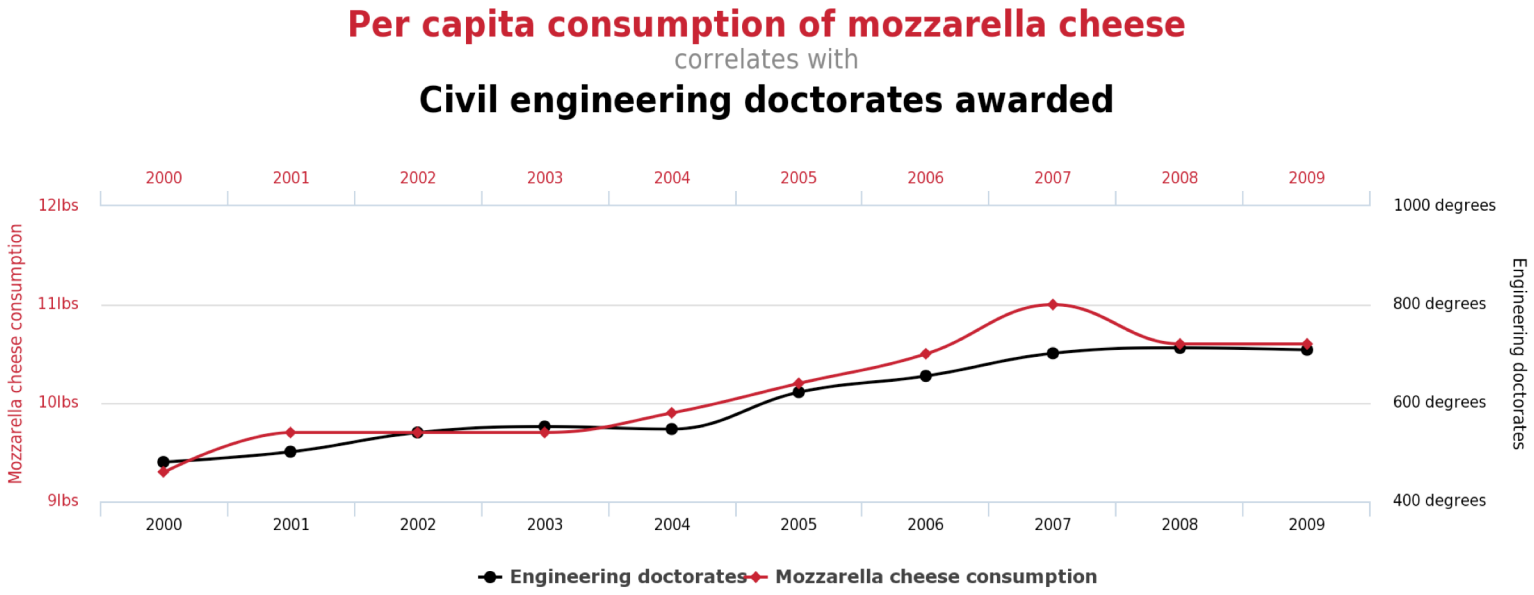
Types of Research Designs

- 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.

Experiments

- (True-)Experiments ultimately aim to support or dismiss hypotheses and models
  - increase understanding of how things work
    - i.e. discover the causality of a phenomenon

- Why do we need experiments to find causality?
  - directionality of correlation problem:
    if X <-> Y, does X -> Y or Y -> X ?



Experiments

- (True-)Experiments ultimately aim to support or dismiss hypotheses and models
  - increase understanding of how things work
    - i.e. discover the causality of a phenomenon

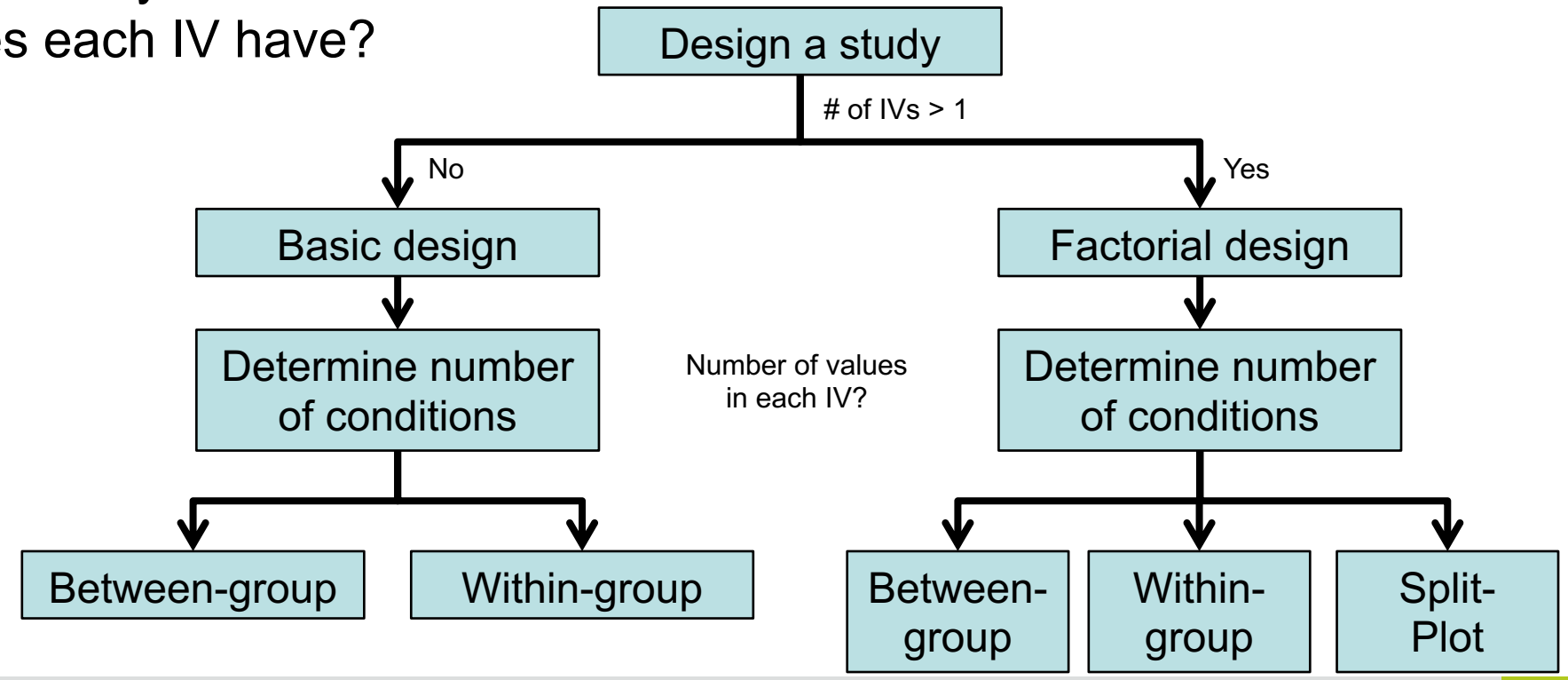
- Why do we need experiments to find causality?
  - directionality of correlation problem:
    if X <-> Y, does X -> Y or Y -> X ?
  - third variable problem:
    if X <-> Y, possibly X<-Z->Y ?

- Only control over all variables can avoid these problems.
  - including unknown variables
  - as far as that is reasonably possible.

Condition Assigment

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 phone unlocking speed when using a numerical PIN, a swipe pattern, or a swipe password”
  - H2: “There is no difference in the time required to locate the privacy settings on a social network site 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 unlock the phone, each participant only does that with one of the three methods
  - 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 participant unlocks the phone with all three methods
- 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

- 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 can be easier to recruit for
    - and sometimes 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!

- It can make sense to try and 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, experts, 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 swipe password 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
    - plan (and enforce) breaks when necessary

Summary Between and Within

- 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
  - − Individual differences can skew results
- Within-groups
  - Each subject does all variants of the experiment
  - \+ Less users required,
  - \+ individual differences canceled out
  - − But often learning effect across conditions is a problem

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 **unlock types** (PIN, swipe pattern, swip password), we also want to investigate the effects of **different tasks** (sitting vs. walking) on unlock speed.
  - 3 x 2 = 6 conditions
    - across the **task dimension**, we can examine the impact of unlock method
    - across the **unlock method 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
  - The three unlock methods are the same speed while sitting
  - PIN and swipe pattern are faster than swipe password while walking

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

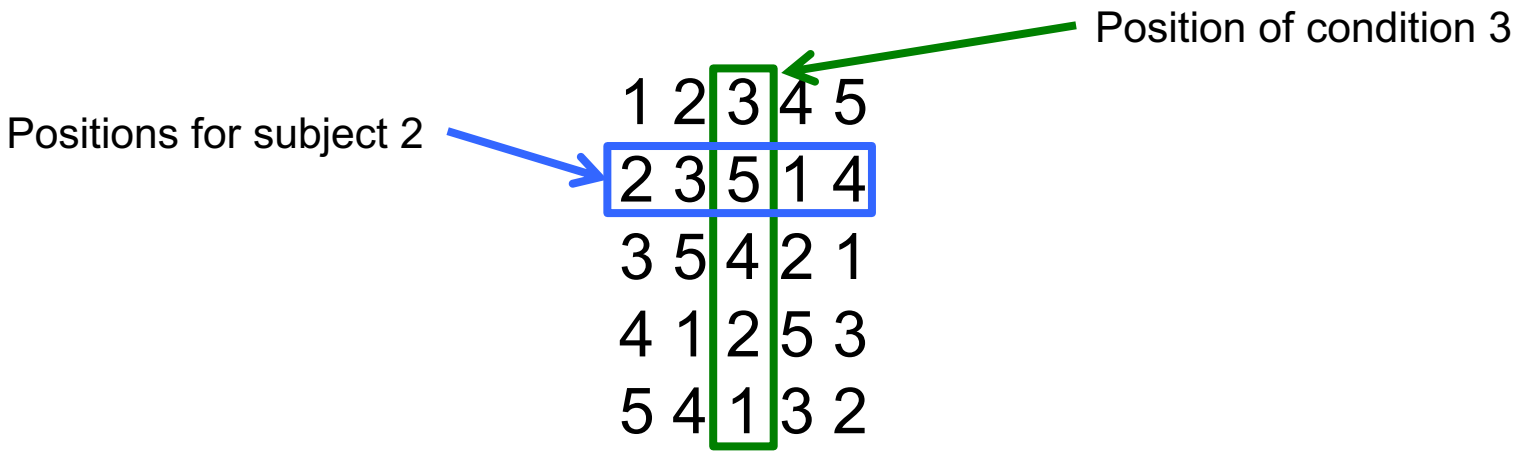
Example Malware Study

3 Independent Variables

- Decompiler (within subjects)
  - Hex-Rays
  - DREAM
  - DREAM++
- Participant (between subjects)
  - Student
  - Expert
- Task (within subjects)
  - Medium
  - Difficult

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.



Study Design