Aalto University

Day 4: Experimental Evaluation

ELEC-D7011 Human Factors Engineering
June 6, 2024
Antti Oulasvirta
Aalto University

SIGN IN
$500,000,000

Q: What evaluation methods are there?

What methods are there?

Analytic evaluation methods

done by interface professionals, no end users necessary

- Usability heuristics
  - *several experts analyze an interface against a handful of principles*

- Walkthroughs
  - *experts and others analyze an interface by considering what a user would have to do a step at a time while performing their task*

What methods are there?

Field studies

requires established end users in their work context

- Ethnography
  - *field worker immerses themselves in a culture to understand what that culture is doing*
- Contextual inquiry
  - *interview methodology that gains knowledge of what people do in their real-world context*

What methods are there?

Self reporting

requires established or potential end users

- interviews
- questionnaires
- surveys

What methods are there?

Modeling

requires detailed interface specifications

- Fitt’s Law
  - *mathematical expression that can predict a user’s time to select a target*
- Keystroke-level model
  - *low-level description of what users would have to do to perform a task that can be used to predict how long it would take them to do it*
- Simulations
  - *Structured models of what users would have to do to perform a task that can also be used to predict time, errors, and effort*

Pros and cons of experiments

How would you evaluate ….

Mobile usability of an LMS (learning management system) app (e.g. MyCourses)

Aalto University

Experimental evaluation

What is an experiment?

To experiment $_{df}$= to cause a change in a phenomenon in order to observe its consequences.

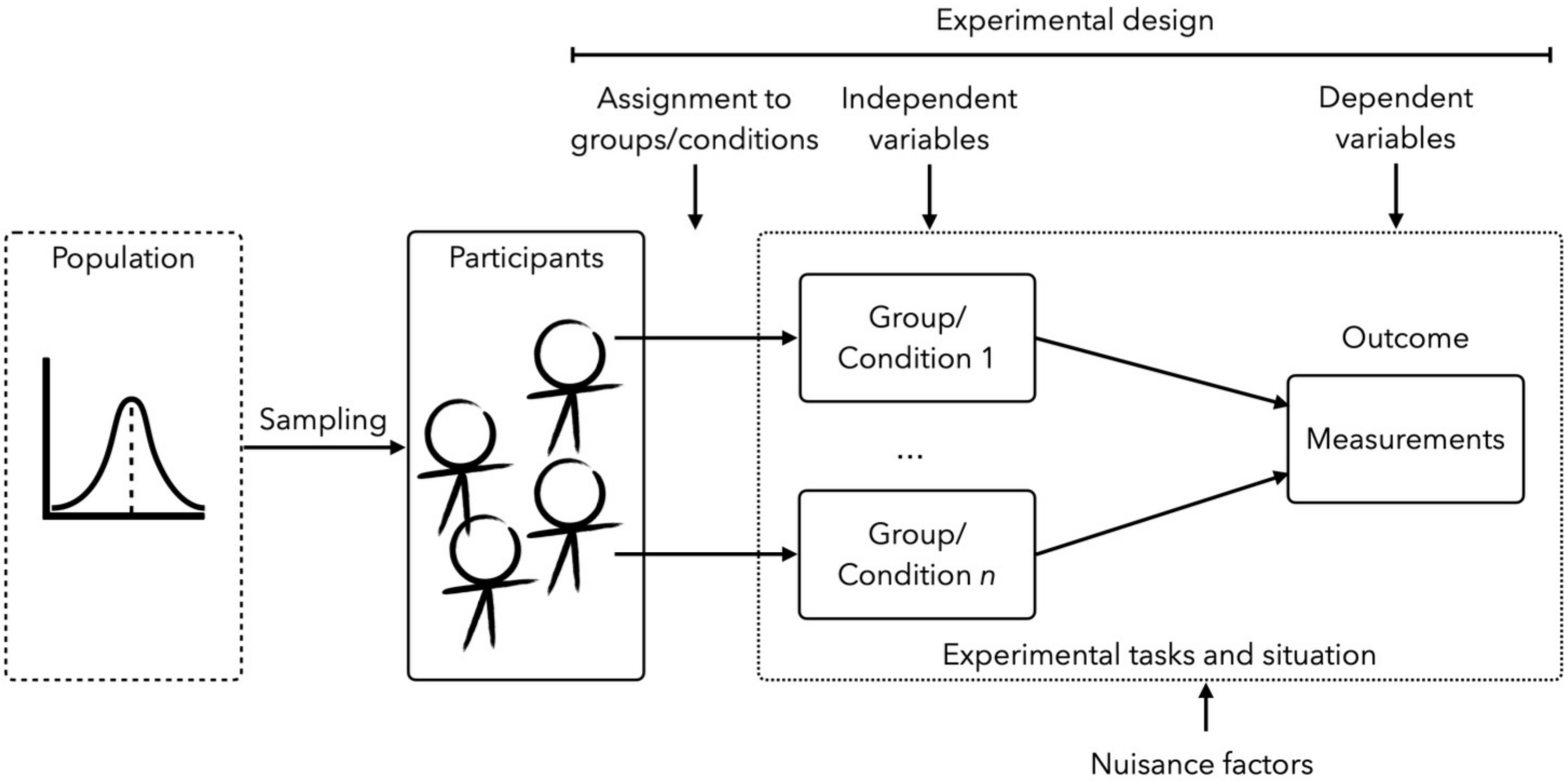
“RIC”: The principle of experimental research

Randomized assignment of subjects to experimental conditions

Independence of observations

Control factors in order to 1) isolate your intervention and 2) eliminate nuisance factors that could produce alternative explanations

Components of an experiment



Components of an experiment

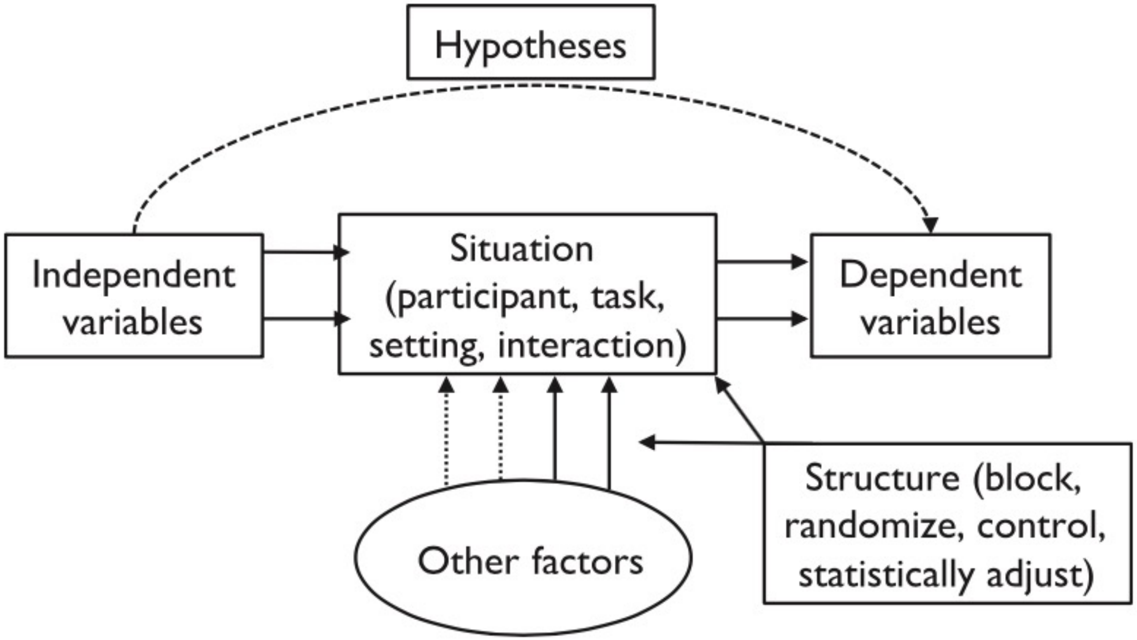

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Hornbaek 2013

Components of an experiment

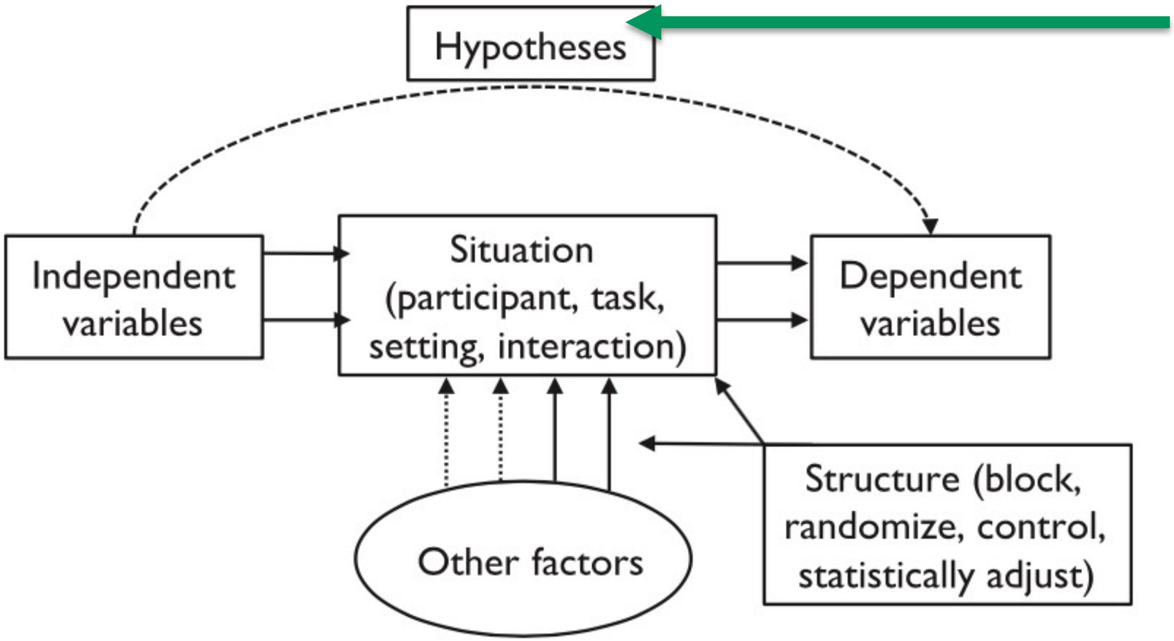
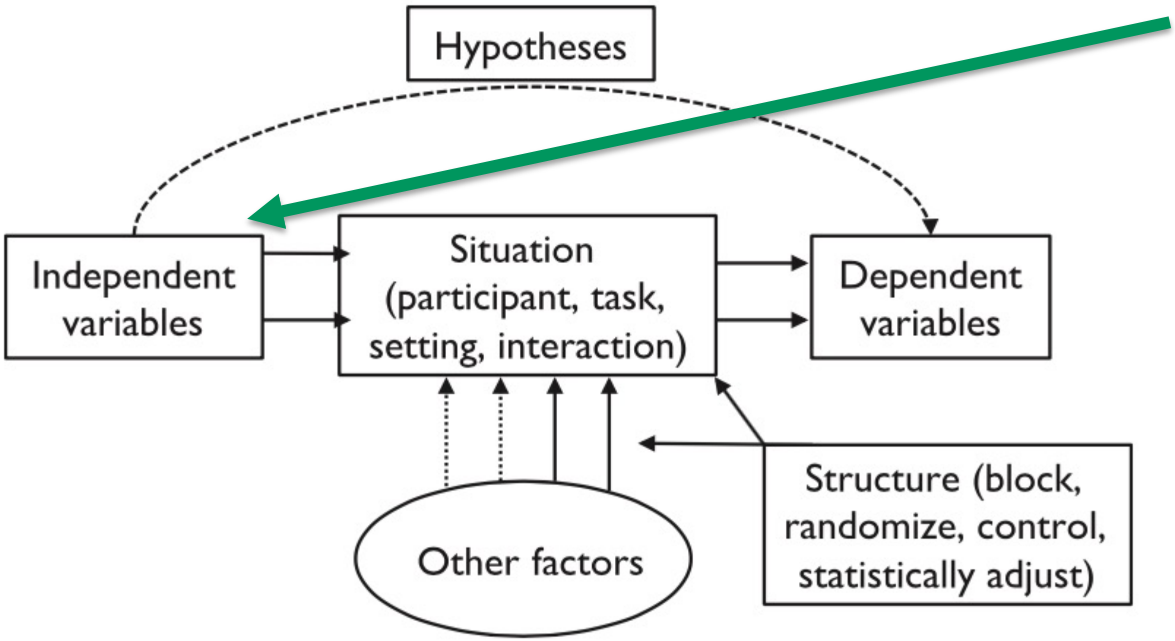

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Research questions that were translated into testable statements that link independent variables to dependent variables

Hornbaek 2013

Components of an experiment

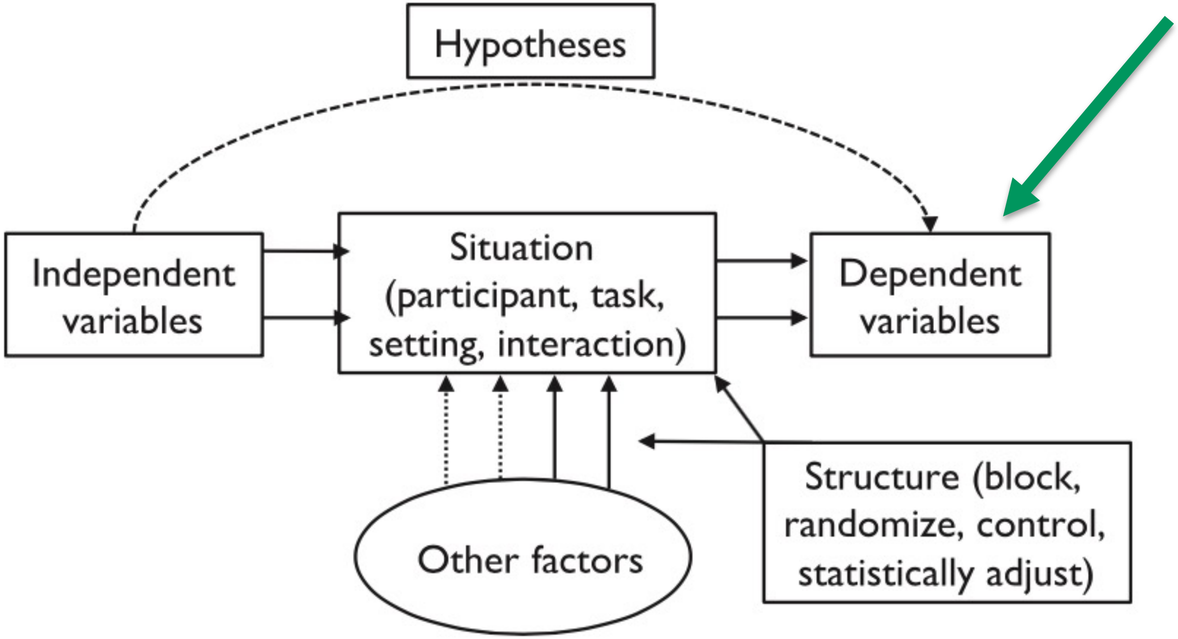

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Manipulated/controlled in the experiment

Fig. 1.1 Typical components of experiments in human–computer interaction.

Hornbaek 2013

Components of an experiment


Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Measured outcome

Fig. 1.1 Typical components of experiments in human–computer interaction.

Hornbaek 2013

Components of an experiment

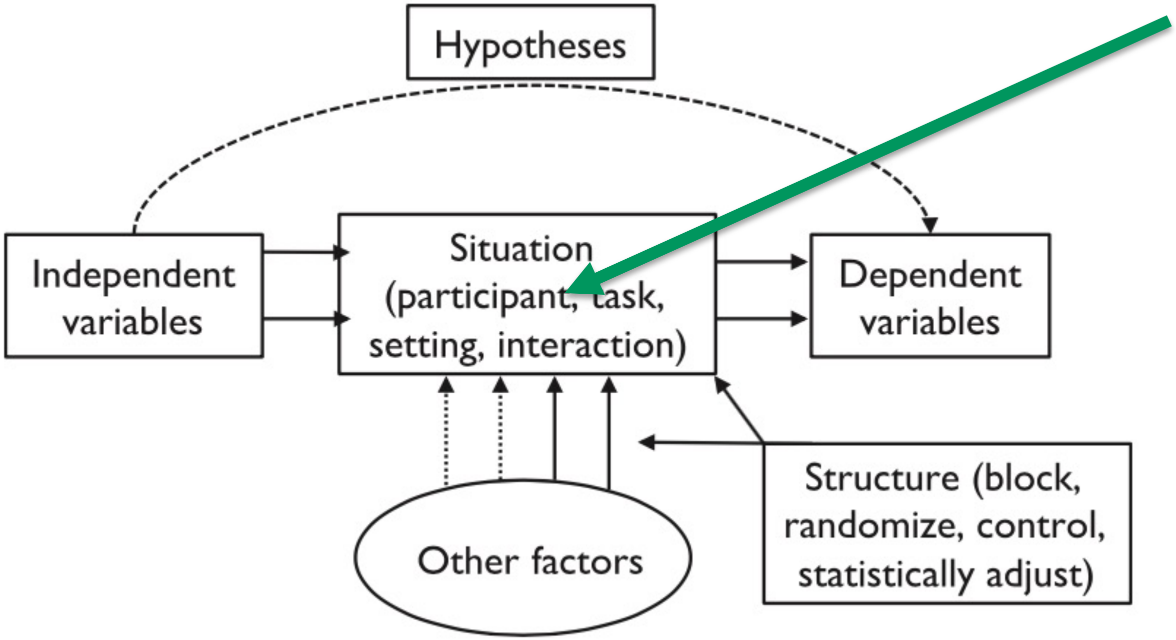
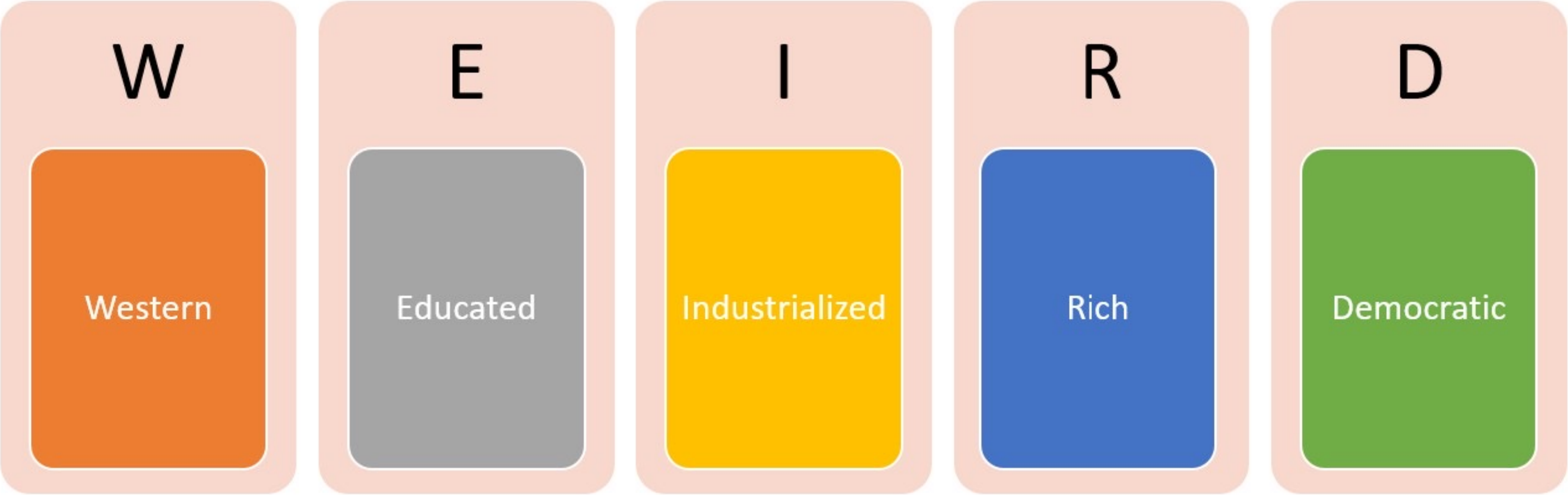

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Participants: volunteers who participate in the experiment, should be representative of the target audience

Hornbaek 2013

WEIRD



W
Western

E
Educated

I
Industrialized

R
Rich

D
Democratic

Components of an experiment

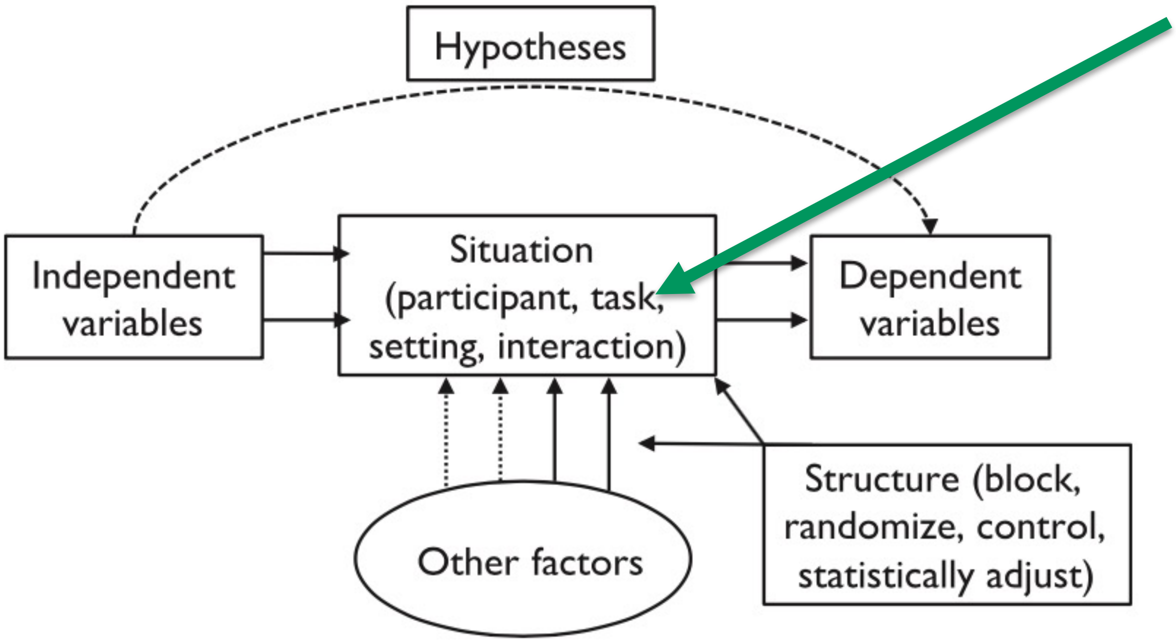

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Experimental task: what we ask participants to carry out in the experiment.

Hornbaek 2013

Components of an experiment

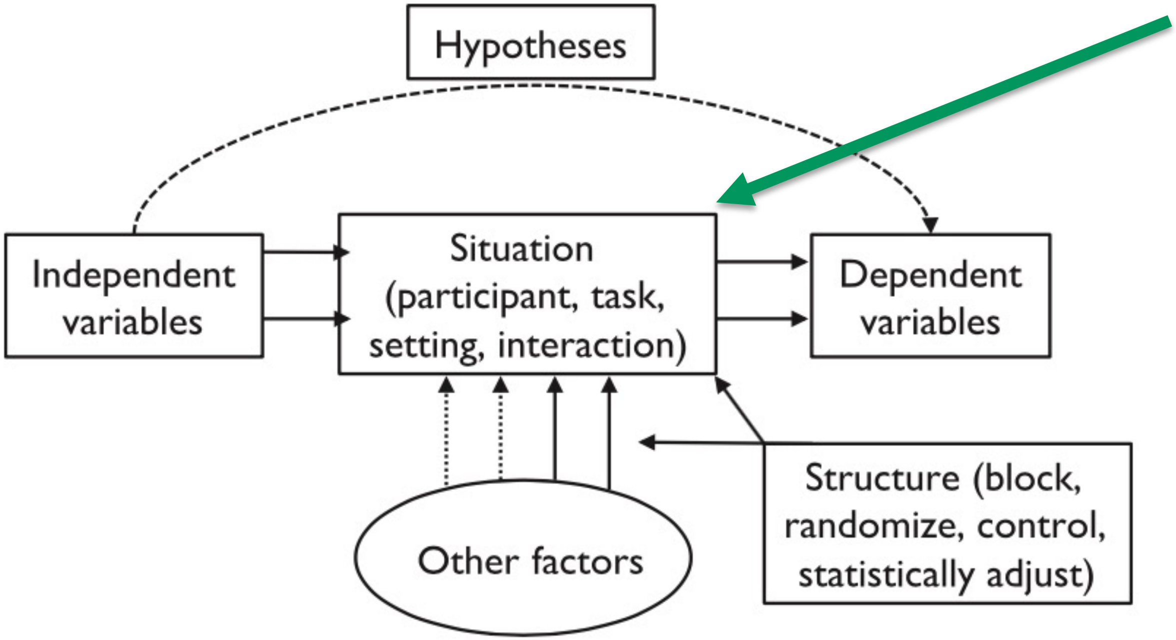

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Condition: the particular setup that the participant is exposed against. Number of conditions depends on the IV.

Hornbaek 2013

Components of an experiment

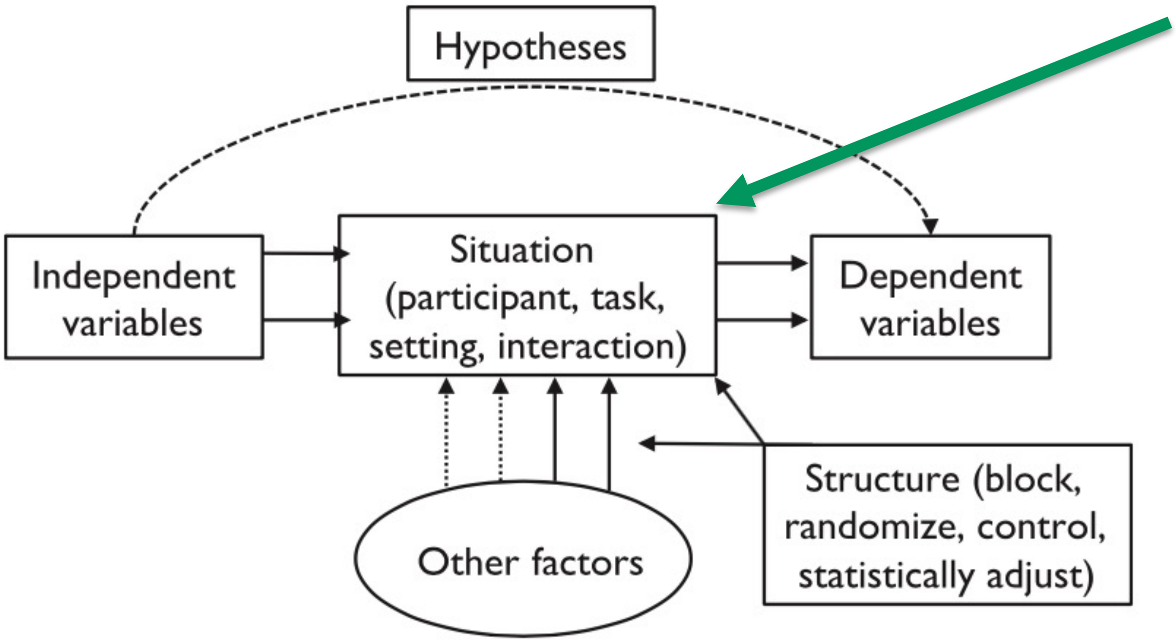

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Condition: the particular setup that the participant is exposed against. Number of conditions depends on the IV.

Hornbaek 2013

Components of an experiment

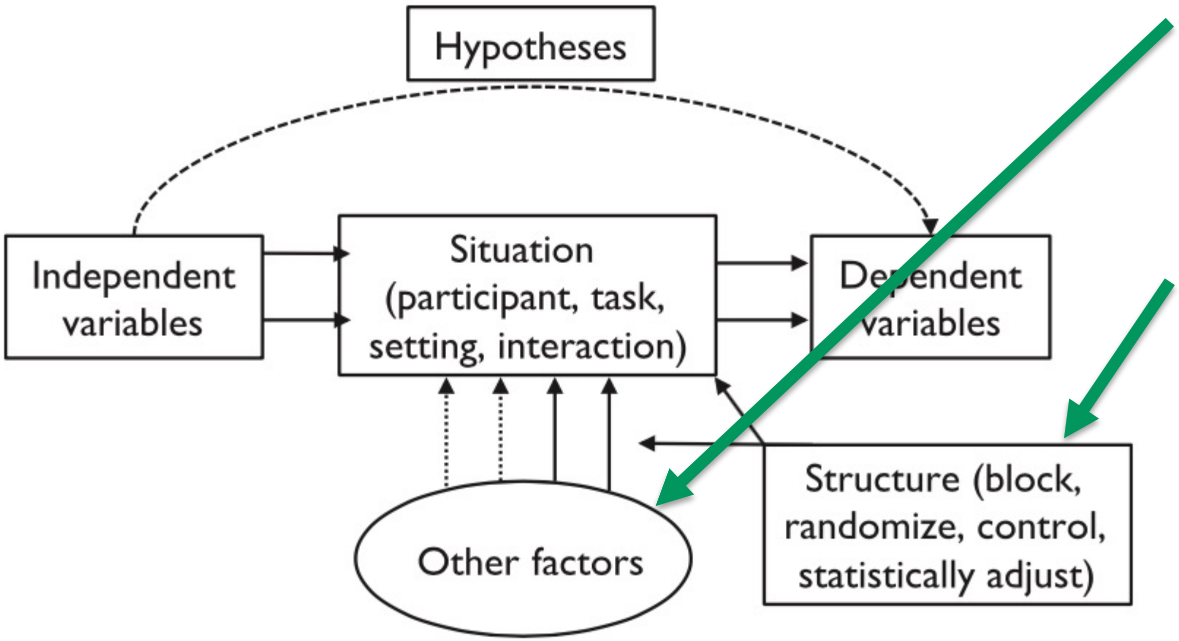

Hypotheses
Independent variables
Situation (participant, task, setting, interaction)
Dependent variables
Other factors
Structure (block, randomize, control, statistically adjust)

Fig. 1.1 Typical components of experiments in human–computer interaction.

Confounding factors/other variables that may affect the outcome

Need to identified and somehow dealt with

Aalto University

Hornbaek 2013

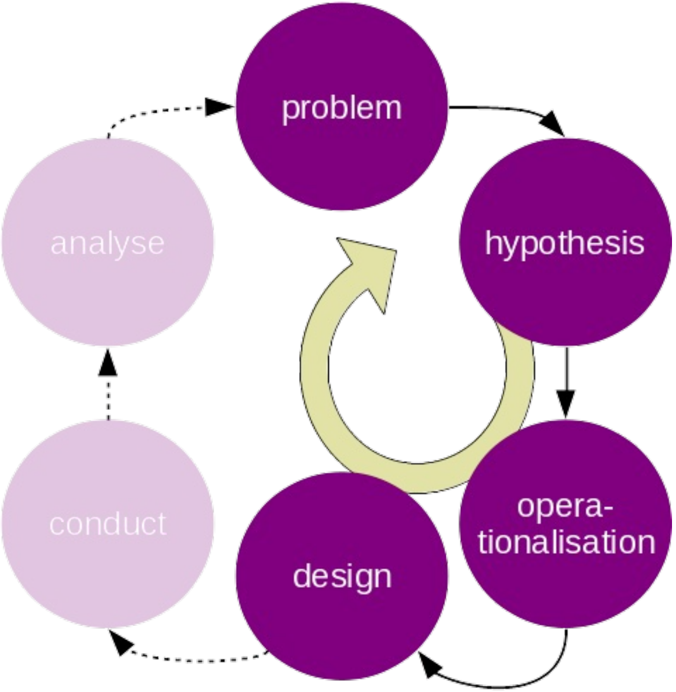
Two most common experimental designs?

Aalto University

How to Run an Experiment

Phases of an Experiment

1. **Research problem**
2. **Hypothesis**
3. **Operationalisation**
   - Independent variables
   - Dependent variables
4. **Study design**
   - Participants
   - Tasks
   - Materials
   - Setting
   - Procedure



Research Questions / Objectives

What do we want to learn from the experiment?

➡ Summarize the reason for the experiment into one question or statement.

Good research objectives are

- **actionable (can be measured somehow)**
- **specific (for a measurement to be determined)**
- **novel (ask questions for which there are no good answer yet)**

E4.A Research questions (7 min)

You want to evaluate the mobile usability of an LMS

Come up with 3 research questions

Typing study: Example questions

What research questions might we ask?

General:

“Does the method increase typing speed?”

“Would users prefer this method over existing ones?”

“Can users type longer with less fatigue?”

User-specific:

“Does the method help elderly people avoid typos?”

“Is this method easier to learn for people new to mobile phones?”

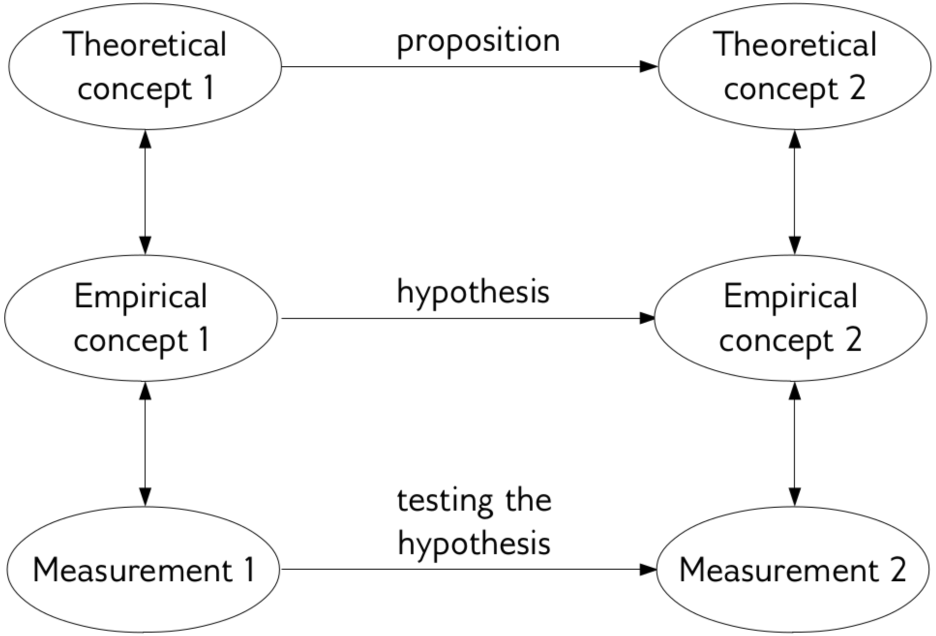
Hypothesis

Proposition

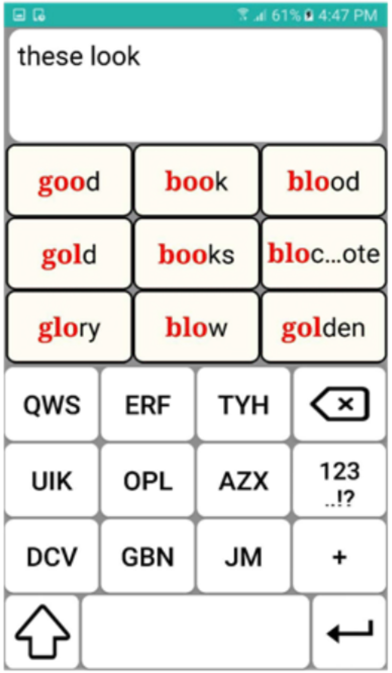
- Is an assumption which connects two theoretical concepts together.

Hypothesis

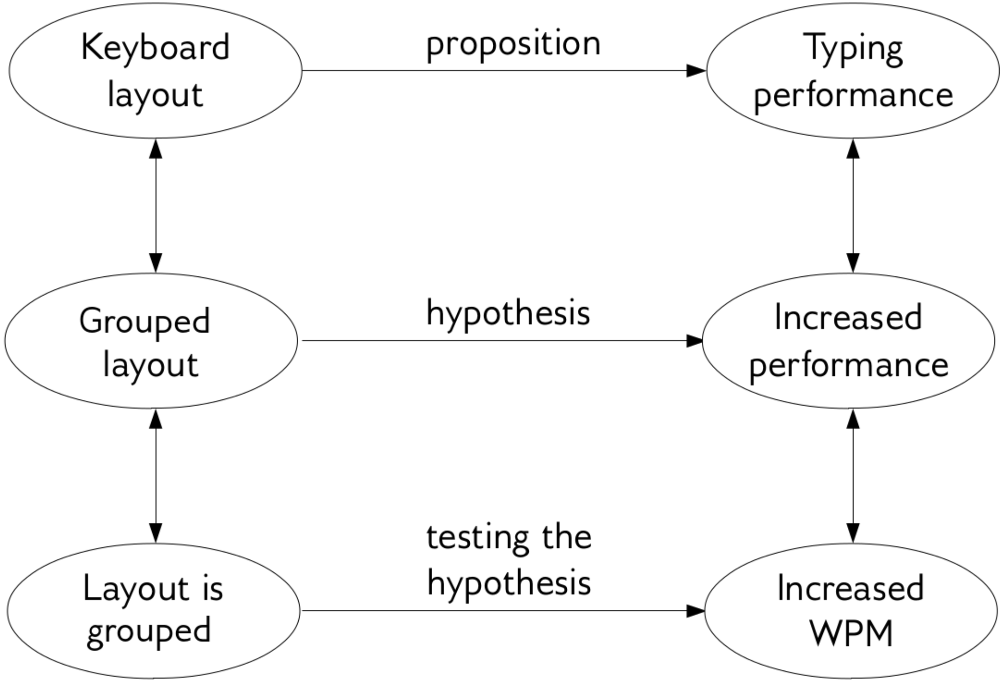
- Is a testable assumption, which connects two empirical concepts together.



Example: Mobile typing



these look



Keyboard layout

proposition

Typing performance

Grouped layout

hypothesis

Increased performance

Layout is grouped

testing the hypothesis

Increased WPM

Operationalisation

- ***Operationalisation* is the translation of a theoretical concept (e.g., typing perforamnce) into an observable or "operable" concept.**

- **The independent variable is manipulated directly by the researcher**
  - Typing method, keyboard layout, typing task etc.
- **The dependent variable is produced in the experiment**
  - WPM, user's rated satisfaction and fatigue, etc.

Typical DVs

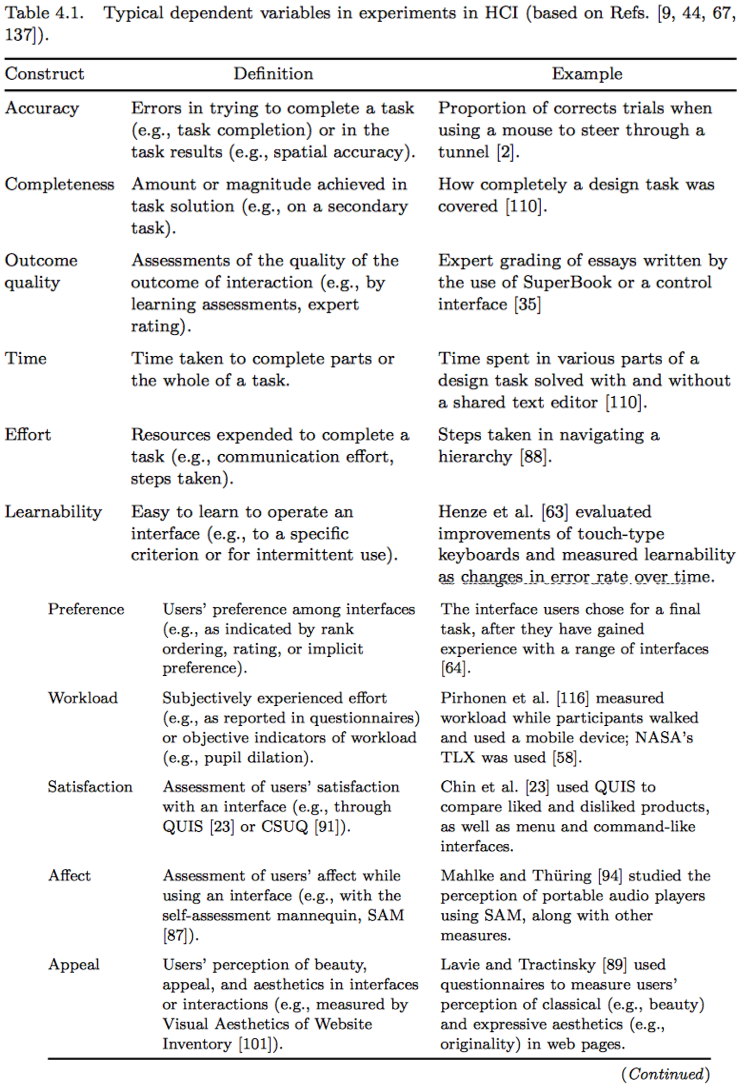
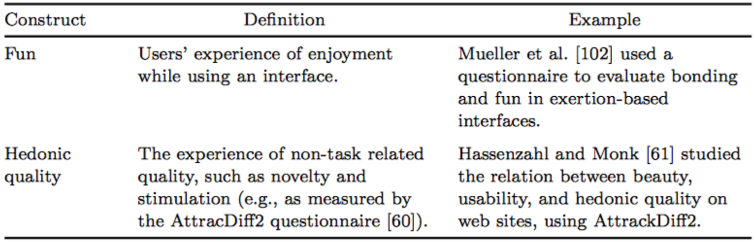


Table 4.1. Typical dependent variables in experiments in HCI (based on Refs. [9, 44, 67, 137]).

Construct	Definition	Example
Accuracy	Errors in trying to complete a task (e.g., task completion) or in the task results (e.g., spatial accuracy).	Proportion of corrects trials when using a mouse to steer through a tunnel [2].
Completeness	Amount or magnitude achieved in task solution (e.g., on a secondary task).	How completely a design task was covered [110].
Outcome quality	Assessments of the quality of the outcome of interaction (e.g., by learning assessments, expert rating).	Expert grading of essays written by the use of SuperBook or a control interface [35]
Time	Time taken to complete parts or the whole of a task.	Time spent in various parts of a design task solved with and without a shared text editor [110].
Effort	Resources expended to complete a task (e.g., communication effort, steps taken).	Steps taken in navigating a hierarchy [88].
Learnability	Easy to learn to operate an interface (e.g., to a specific criterion or for intermittent use).	Henze et al. [63] evaluated improvements of touch-type keyboards and measured learnability as changes in error rate over time.
Preference	Users' preference among interfaces (e.g., as indicated by rank ordering, rating, or implicit preference).	The interface users chose for a final task, after they have gained experience with a range of interfaces [64].
Workload	Subjectively experienced effort (e.g., as reported in questionnaires) or objective indicators of workload (e.g., pupil dilation).	Pirhonen et al. [116] measured workload while participants walked and used a mobile device; NASA's TLX was used [58].
Satisfaction	Assessment of users' satisfaction with an interface (e.g., through QUIS [23] or CSUQ [91]).	Chin et al. [23] used QUIS to compare liked and disliked products, as well as menu and command-like interfaces.
Affect	Assessment of users' affect while using an interface (e.g., with the self-assessment mannequin, SAM [87]).	Mahlke and Thüring [94] studied the perception of portable audio players using SAM, along with other measures.
Appeal	Users' perception of beauty, appeal, and aesthetics in interfaces or interactions (e.g., measured by Visual Aesthetics of Website Inventory [101]).	Lavie and Tractinsky [89] used questionnaires to measure users' perception of classical (e.g., beauty) and expressive aesthetics (e.g., originality) in web pages.

(*Continued*)



Construct	Definition	Example
Fun	Users' experience of enjoyment while using an interface.	Mueller et al. [102] used a questionnaire to evaluate bonding and fun in exertion-based interfaces.
Hedonic quality	The experience of non-task related quality, such as novelty and stimulation (e.g., as measured by the AttracDiff2 questionnaire [60]).	Hassenzahl and Monk [61] studied the relation between beauty, usability, and hedonic quality on web sites, using AttrackDiff2.

Check if there are standard measures and build on prior work.

Hornbaek 2013

E4.B Research questions (5 min)

Write down research hypotheses

1 - Translate your research questions (1-2) into experimental hypothesis.

2 - Which independent and dependent variables are suitable to test your hypothesis?

Participants

Should be representative of your target audience

- Students, children, professionals, etc.
- Expert users, novice users
- Demographics
- Reasonable amount

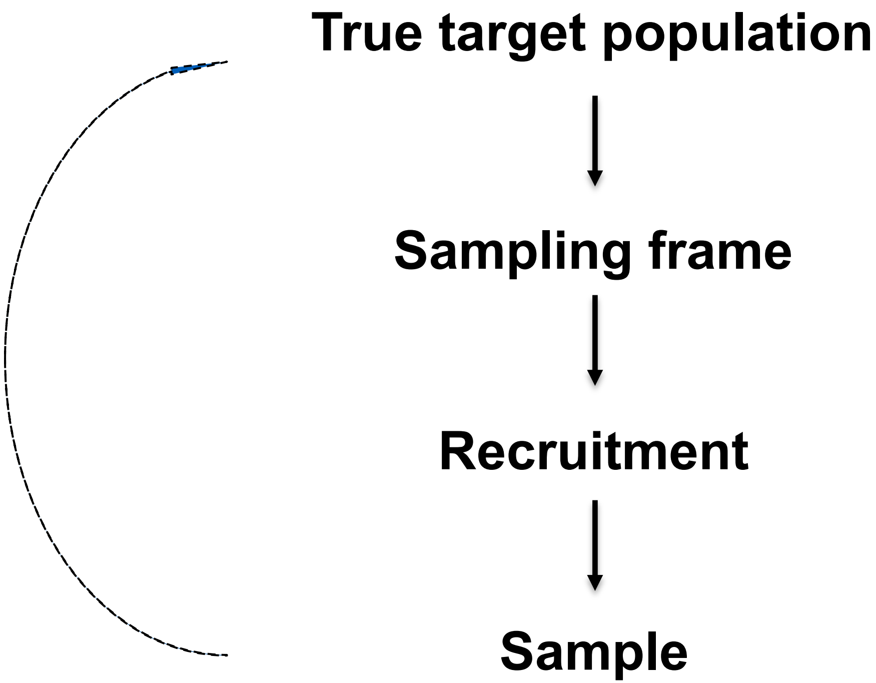
Participants’ motivation can affect recruitment and results

- Rewards: Money, food, coupons, (if any)
- Beware of selection bias

Recruitment

- Posters, email, snowball sampling, ads, web, etc.

Sampling frame



True target population	“Students of Aalto”
Sampling frame	“Students at Väre”
Recruitment	“5 out of 13 students we asked agreed to join when we asked on Friday evening”
Sample	The 5 students

How to Treat Participants

Don’t waste their time

Informed consent

Privacy, Confidentiality, Anonymity

Make them feel comfortable

Estimate risks

Tasks

- **Should be realistic**
  - Represent tasks users would actually carry out in “real life”
  - While sufficiently abstract to reduce or handle confounding factors so that it is possible to infer causality
- **Should be efficient**
  - Think: Number of observations you make per unit of time
- **Should be controllable**
  - We can eliminate nuisance factors
  - We can minimize unnecessary variance

Example: Typing:

Task A: type a sequence of random words shown on the screen – not representative

Task B: type a message to your friend – very realistic, but we can’t control it

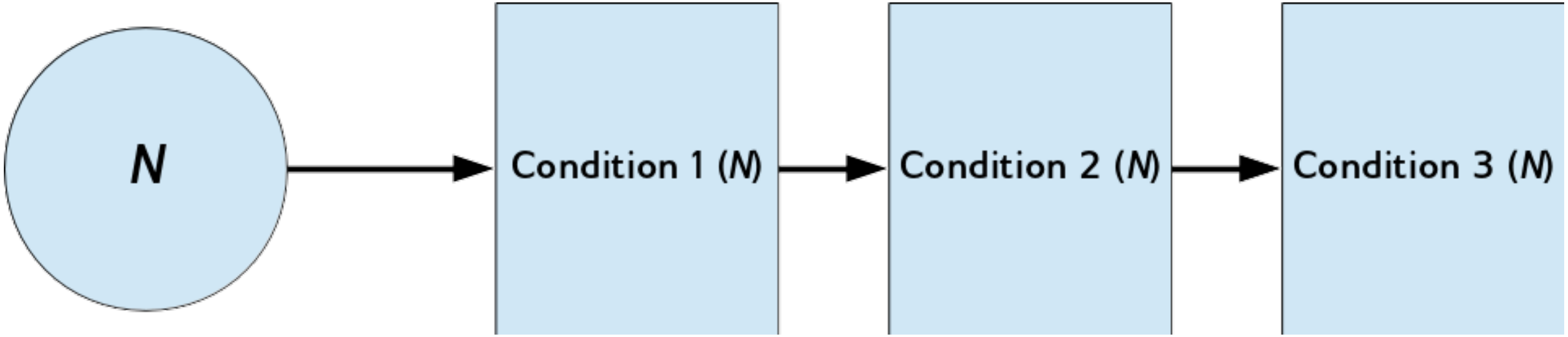
Task C: type a sentence given on the screen – balances realism and control

Basic experimental designs

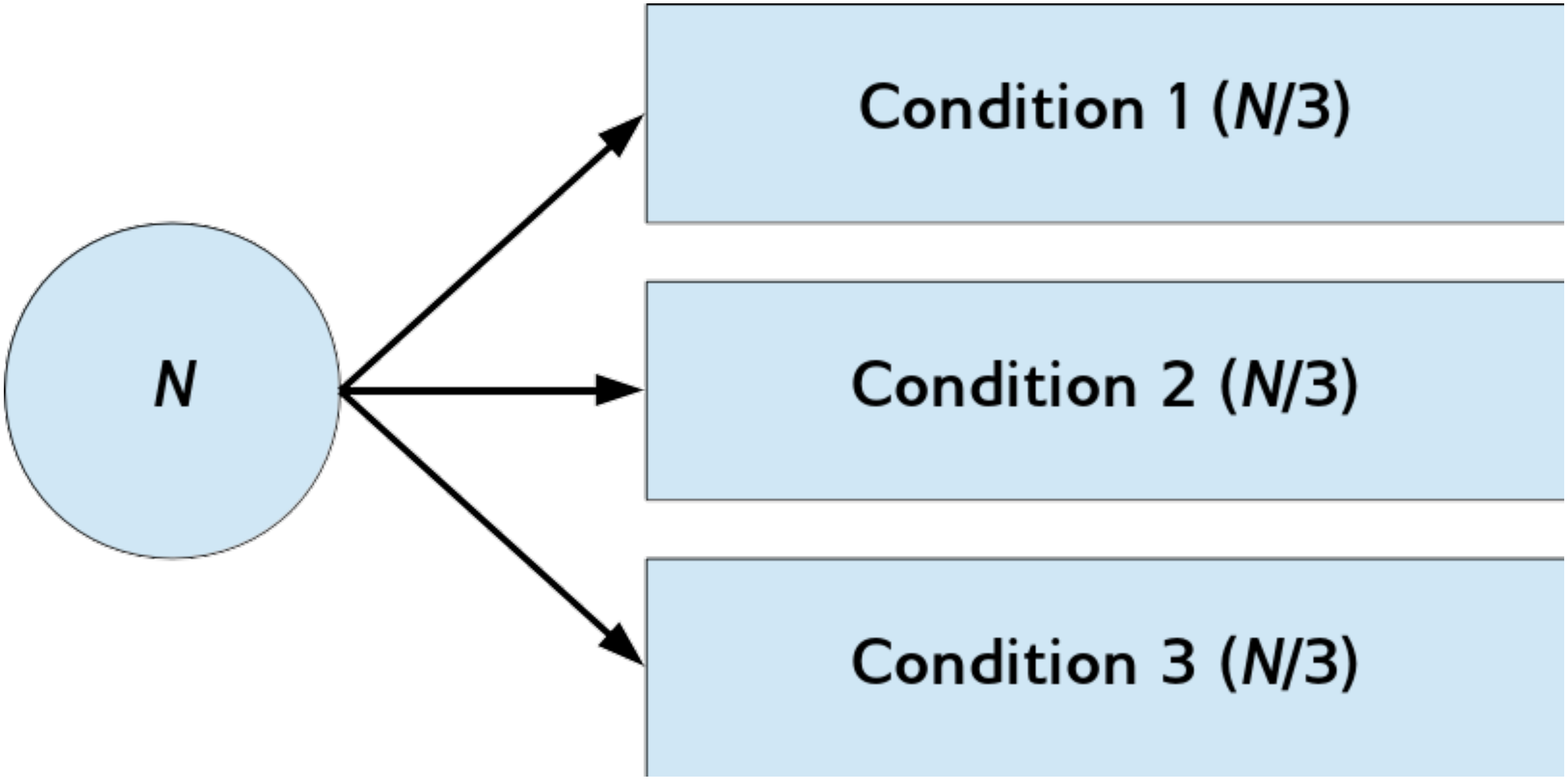
Between-subjects design
Within-subject design

Completely randomized factorial design
Latin square design
One group posttest-only design
Fixed effect model
Random effects model
Interrupted time series experiment
Dependent samples t-test
…

Within-Subjects Design



Between-Subject Design



Q: Benefits and drawbacks of the two?

Conducting and Experiment

Advice for treating each participant the same

- Practice the procedure
- Once you start collecting data, don't make changes
- If several experimenters, agree on a script

Data collection methods

Live notetaking

Video capture (face, body)

Screen capture

Audio capture

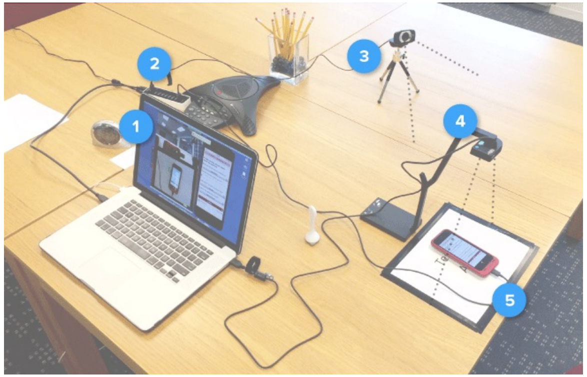
Physiological measurements

Logging

- Task performance (accuracy, errors, time)
- Task trajectories (how a task was done)

Think aloud protocols

- Asking the participant to talk aloud what he/she is thinking



E4.C Dissecting an experimental paper (10 min)

https://journals.sagepub.com/doi/full/10.1177/0018720819891291

Identify and report (use bullets):

- ***Hypothesis***
- ***IVs***
- ***DVs***
- ***Participants (sampling frame?)***
- ***Materials***
- ***Tasks***

Reporting an Experiment

Method section

- Others need to know everything that might affect your results
- Guideline APA style: https://www.scribbr.com/apa-style/methods-section/

Aalto University

Validity

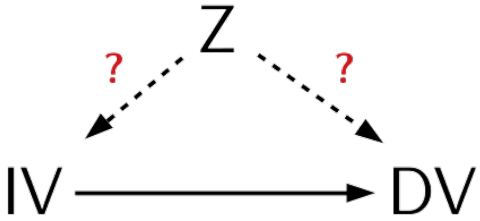
What makes an experiment good or bad

✋ What does ‘validity’ mean?

William R.. Shadish, Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning.

Internal Validity

- Internal Validity is the extend to which causal claims are justified.
- Would the observed differences in the DV be present without variation in the IV?
- Are we measuring the right thing?
- Is there anything else that could explain the effect?



External Validity

- External validity is the extent to which the inferences from the study generalize.
- Do the data offer enough evidence to support the hypothesis such that the result can be generalized beyond the study?

“Threat to validity” (aka nuisance factor)

= Anything that can go wrong

= Anything that threatens your ability to draw solid conclusions

Example: Usability evaluation (next)

In experimental research, established taxonomies for threats

- (see next slide)

Typical threats to validity in usability testing

Random incidents

Technical problems

Learning effect

Fatigue

Drop-out

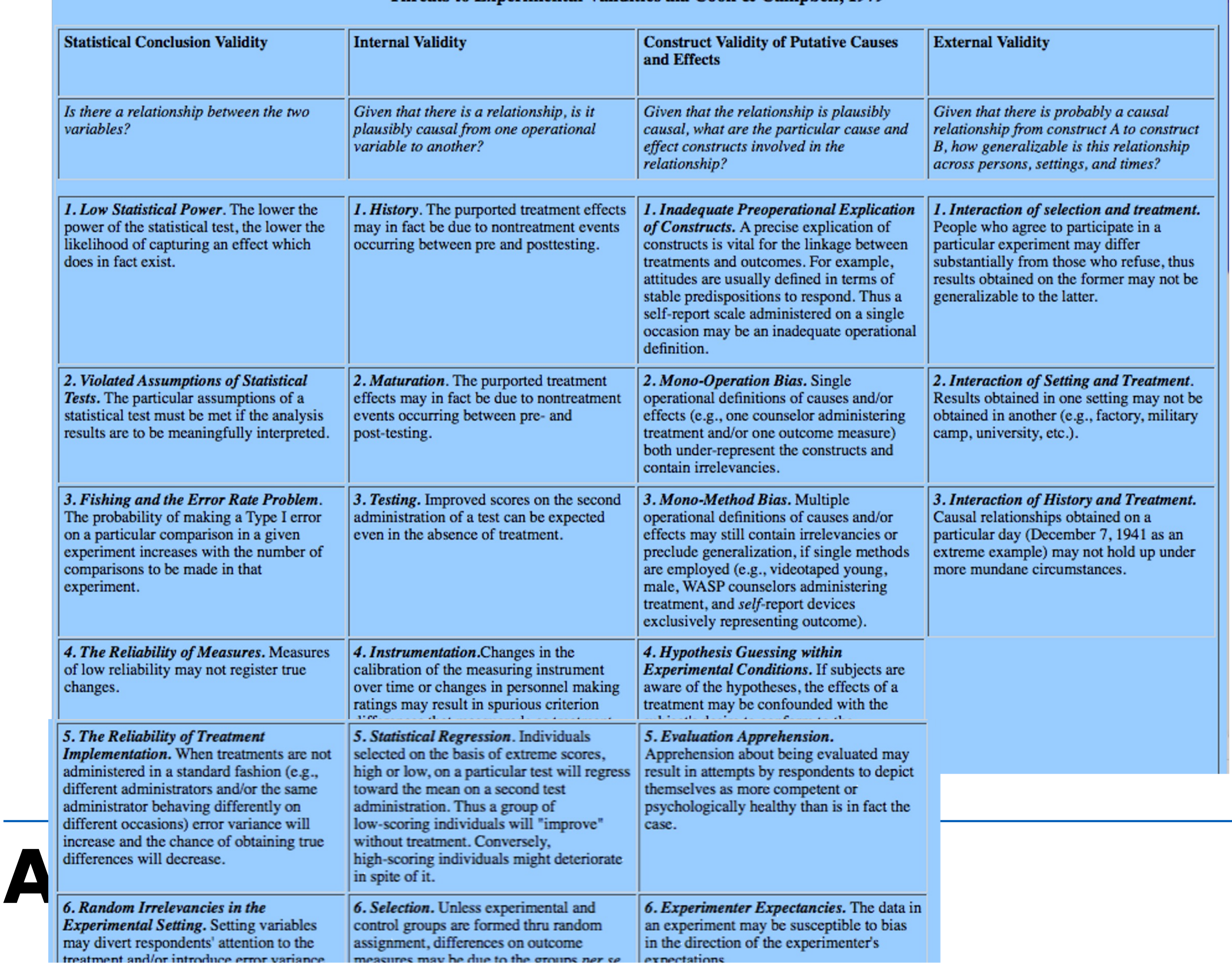
Self-selection

Guessing of the experiment’s purpose

Evaluator effect

These are risks that may or may not (if you’re lucky) hamper your conclusions

By careful design of an experiment, you can eliminate or mitigate these!



Threats to Experimental Validities after Cook & Campbell, 1979

Statistical Conclusion Validity	Internal Validity	Construct Validity of Putative Causes and Effects	External Validity
Is there a relationship between the two variables?	*Given that there is a relationship, is it plausibly causal from one operational variable to another?*	*Given that the relationship is plausibly causal, what are the particular cause and effect constructs involved in the relationship?*	*Given that there is probably a causal relationship from construct A to construct B, how generalizable is this relationship across persons, settings, and times?*
1. Low Statistical Power. The lower the power of the statistical test, the lower the likelihood of capturing an effect which does in fact exist.	***1. History***. The purported treatment effects may in fact be due to nontreatment events occurring between pre and posttesting.	***1. Inadequate Preoperational Explication of Constructs.*** A precise explication of constructs is vital for the linkage between treatments and outcomes. For example, attitudes are usually defined in terms of stable predispositions to respond. Thus a self-report scale administered on a single occasion may be an inadequate operational definition.	***1. Interaction of selection and treatment.*** People who agree to participate in a particular experiment may differ substantially from those who refuse, thus results obtained on the former may not be generalizable to the latter.
2. Violated Assumptions of Statistical Tests. The particular assumptions of a statistical test must be met if the analysis results are to be meaningfully interpreted.	***2. Maturation***. The purported treatment effects may in fact be due to nontreatment events occurring between pre- and post-testing.	***2. Mono-Operation Bias.*** Single operational definitions of causes and/or effects (e.g., one counselor administering treatment and/or one outcome measure) both under-represent the constructs and contain irrelevancies.	***2. Interaction of Setting and Treatment***. Results obtained in one setting may not be obtained in another (e.g., factory, military camp, university, etc.).
3. Fishing and the Error Rate Problem. The probability of making a Type I error on a particular comparison in a given experiment increases with the number of comparisons to be made in that experiment.	***3. Testing.*** Improved scores on the second administration of a test can be expected even in the absence of treatment.	***3. Mono-Method Bias.*** Multiple operational definitions of causes and/or effects may still contain irrelevancies or preclude generalization, if single methods are employed (e.g., videotaped young, male, WASP counselors administering treatment, and *self*-report devices exclusively representing outcome).	***3. Interaction of History and Treatment.*** Causal relationships obtained on a particular day (December 7, 1941 as an extreme example) may not hold up under more mundane circumstances.
4. The Reliability of Measures. Measures of low reliability may not register true changes.	***4. Instrumentation.***Changes in the calibration of the measuring instrument over time or changes in personnel making ratings may result in spurious criterion differences that masquerade as treatment	***4. Hypothesis Guessing within Experimental Conditions.*** If subjects are aware of the hypotheses, the effects of a treatment may be confounded with the subject's desire to conform to the	
5. The Reliability of Treatment Implementation. When treatments are not administered in a standard fashion (e.g., different administrators and/or the same administrator behaving differently on different occasions) error variance will increase and the chance of obtaining true differences will decrease.	***5. Statistical Regression***. Individuals selected on the basis of extreme scores, high or low, on a particular test will regress toward the mean on a second test administration. Thus a group of low-scoring individuals will "improve" without treatment. Conversely, high-scoring individuals might deteriorate in spite of it.	***5. Evaluation Apprehension.*** Apprehension about being evaluated may result in attempts by respondents to depict themselves as more competent or psychologically healthy than is in fact the case.	
6. Random Irrelevancies in the Experimental Setting. Setting variables may divert respondents' attention to the treatment and/or introduce error variance	***6. Selection.*** Unless experimental and control groups are formed thru random assignment, differences on outcome measures may be due to the groups *per se*	***6. Experimenter Expectancies.*** The data in an experiment may be susceptible to bias in the direction of the experimenter's expectations.	

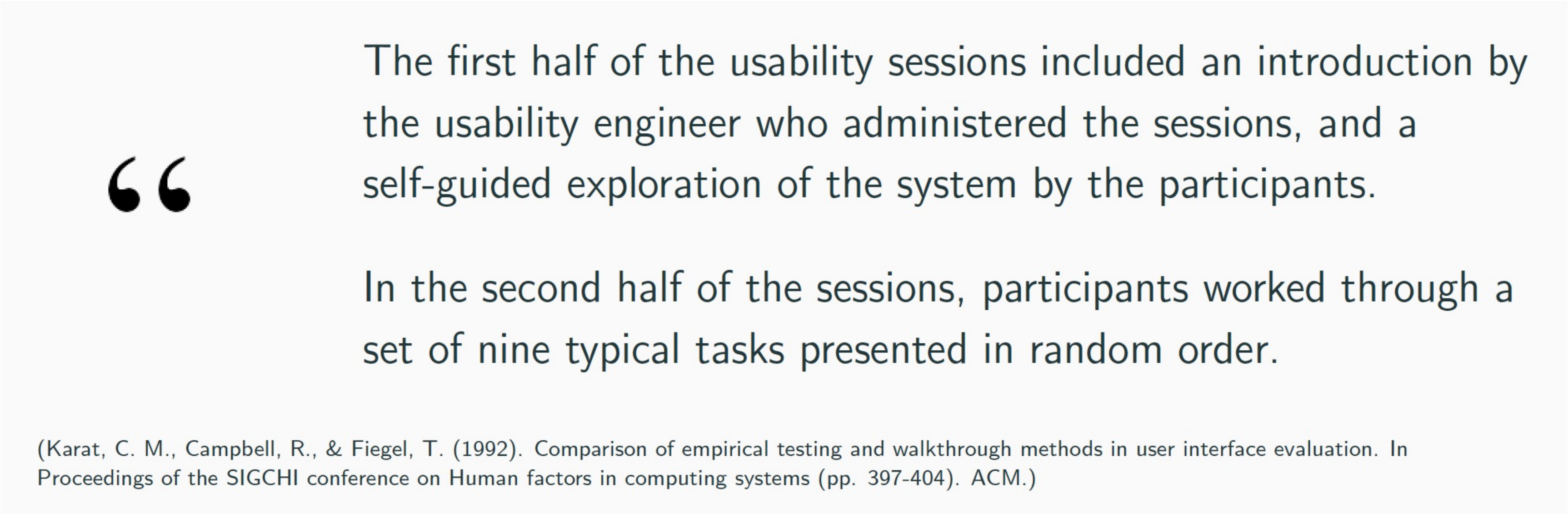
A

Mitigating nuisance variables

1. Direct intervention
2. Prescreening
3. Keep constant for everybody
4. Vary systematically, include in the experimental design as an IV
5. Randomize
6. Post hoc analysis

Common techniques
Single-blind, Double-blind
Deception, Disguised experiment
Unrelated-experiment technique, Postexperimental debriefing
Multiple researchers, Experimenter-expectancy control groups

What went wrong?



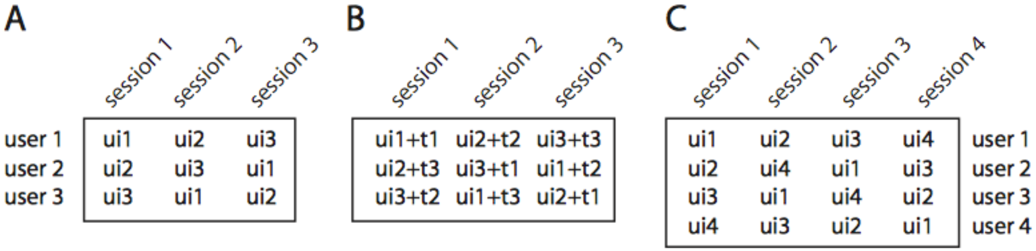
“ The first half of the usability sessions included an introduction by the usability engineer who administered the sessions, and a self-guided exploration of the system by the participants.

In the second half of the sessions, participants worked through a set of nine typical tasks presented in random order.

(Karat, C. M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 397-404). ACM.)

Counterbalancing

How to get rid of a “carry over effect” like learning the UI or getting tired?



A

	session 1	session 2	session 3
user 1	ui1	ui2	ui3
user 2	ui2	ui3	ui1
user 3	ui3	ui1	ui2

B

session 1	session 2	session 3
ui1+t1	ui2+t2	ui3+t3
ui2+t3	ui3+t1	ui1+t2
ui3+t2	ui1+t3	ui2+t1

C

session 1	session 2	session 3	session 4	
ui1	ui2	ui3	ui4	user 1
ui2	ui4	ui1	ui3	user 2
ui3	ui1	ui4	ui2	user 3
ui4	ui3	ui2	ui1	user 4

Fig. 4.2 Counterbalancing with Latin and Greco-Latin Squares. Panel A shows a within-subjects design for three user interfaces (ui1–ui3), each used in a sequence of sessions (session 1–session 3) by participants (user 1–user 3). Panel B shows a Greco–Latin Square that crosses user interfaces with tasks (t1–t3). Panel C shows a 4 $*$ 4 Latin square, for a situation with four user interfaces (or four other levels of an independent variable). That square is balanced for first-order effects in that each user interface is followed in the next session by any other user interface the same number of times.

Hornbaek 2013

General guidelines for experiments

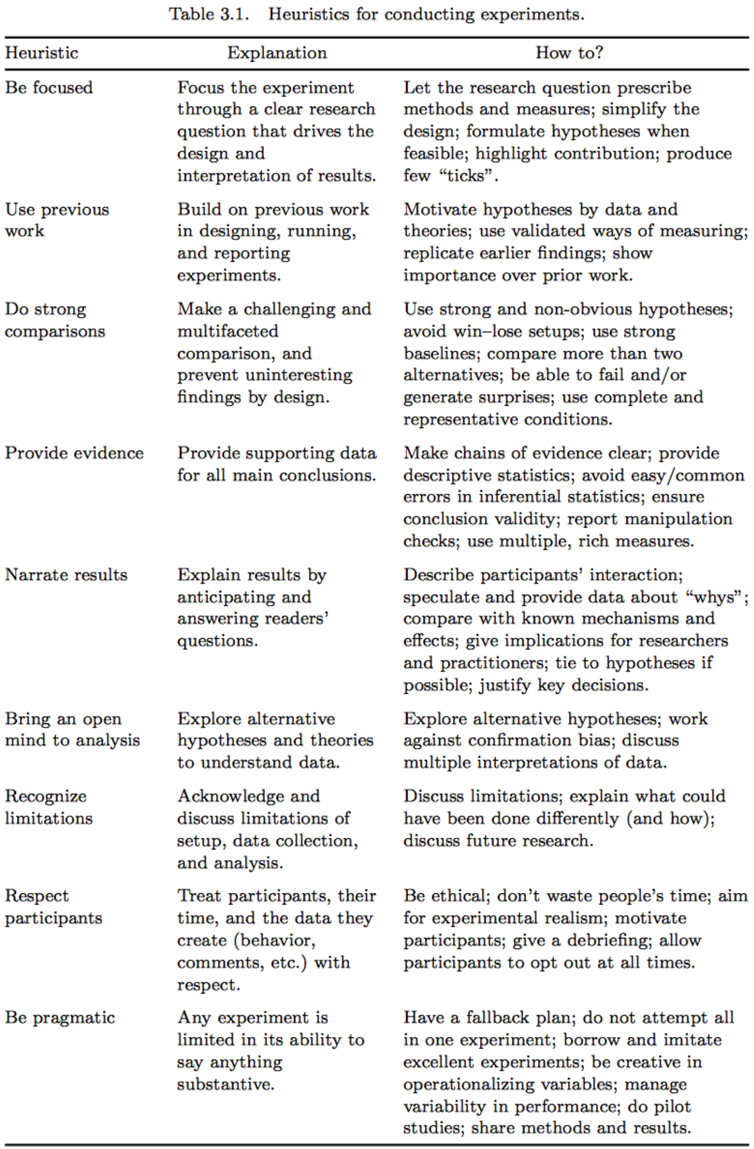


Table 3.1. Heuristics for conducting experiments.

Heuristic	Explanation	How to?
Be focused	Focus the experiment through a clear research question that drives the design and interpretation of results.	Let the research question prescribe methods and measures; simplify the design; formulate hypotheses when feasible; highlight contribution; produce few "ticks".
Use previous work	Build on previous work in designing, running, and reporting experiments.	Motivate hypotheses by data and theories; use validated ways of measuring; replicate earlier findings; show importance over prior work.
Do strong comparisons	Make a challenging and multifaceted comparison, and prevent uninteresting findings by design.	Use strong and non-obvious hypotheses; avoid win–lose setups; use strong baselines; compare more than two alternatives; be able to fail and/or generate surprises; use complete and representative conditions.
Provide evidence	Provide supporting data for all main conclusions.	Make chains of evidence clear; provide descriptive statistics; avoid easy/common errors in inferential statistics; ensure conclusion validity; report manipulation checks; use multiple, rich measures.
Narrate results	Explain results by anticipating and answering readers' questions.	Describe participants' interaction; speculate and provide data about "whys"; compare with known mechanisms and effects; give implications for researchers and practitioners; tie to hypotheses if possible; justify key decisions.
Bring an open mind to analysis	Explore alternative hypotheses and theories to understand data.	Explore alternative hypotheses; work against confirmation bias; discuss multiple interpretations of data.
Recognize limitations	Acknowledge and discuss limitations of setup, data collection, and analysis.	Discuss limitations; explain what could have been done differently (and how); discuss future research.
Respect participants	Treat participants, their time, and the data they create (behavior, comments, etc.) with respect.	Be ethical; don't waste people's time; aim for experimental realism; motivate participants; give a debriefing; allow participants to opt out at all times.
Be pragmatic	Any experiment is limited in its ability to say anything substantive.	Have a fallback plan; do not attempt all in one experiment; borrow and imitate excellent experiments; be creative in operationalizing variables; manage variability in performance; do pilot studies; share methods and results.

Hornbaek 2013

E4.D Critiquing an experiment (10 min)

https://journals.sagepub.com/doi/full/10.1177/0018720819891291

Identify relevant threats to validity
Use the technical terms by Cook & Campbell

Aalto University

Pairwork topics

By far the hardest task so far on this course!

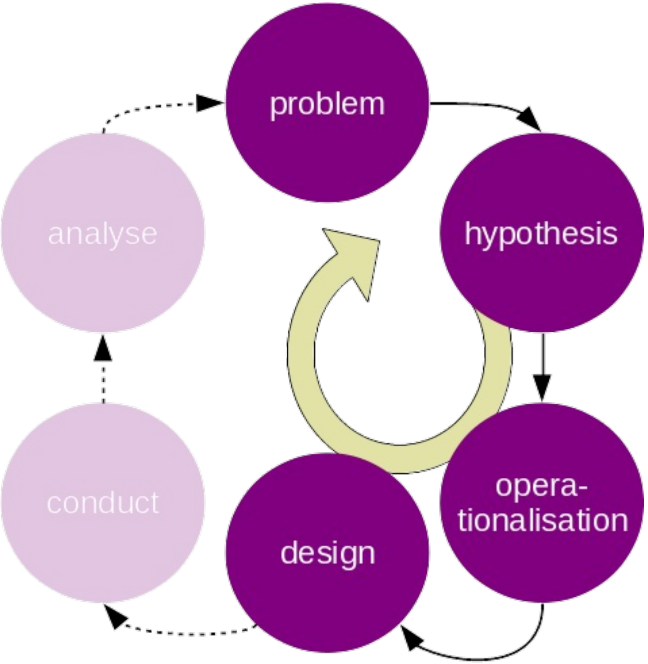
Evaluate a mobile user interface

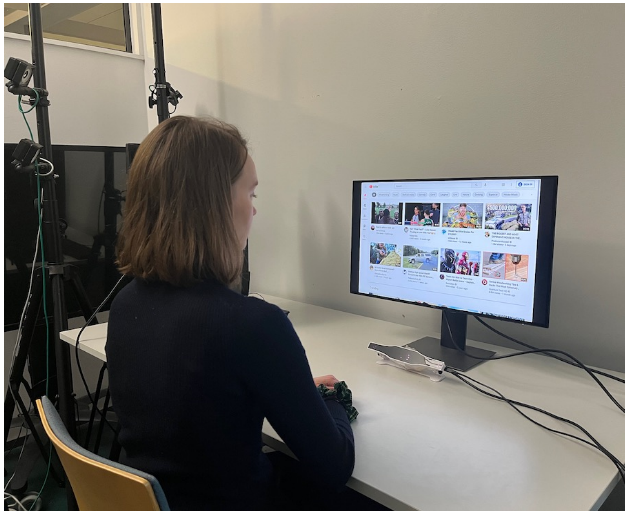
Your goal is to evaluate the *mobile usability* of Aalto MyCourses

Design an experiment and pilot it with 1 user

Technical pilot (N=1)

The purpose is to test the experimental design itself





Do the tasks, instructions, measurements etc. work?

Instructions and presentation

Design an experiment (ideal)

- could be run in ~40 hours of work
- budget of max. 5000 eur

Consider the following

- Hypotheses
- Participants
- Experimental design
- Materials and Tasks
- Procedure
- Measurements

Run a "technical pilot" (N=1)

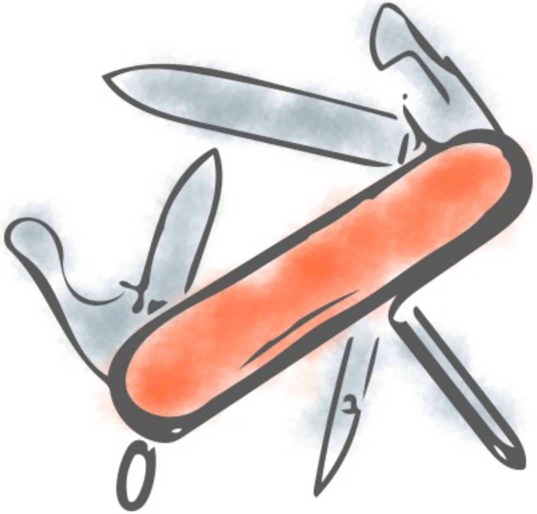
- Anyone on campus who is not you
- Use proxy measurements
- Report what you can

Presentation slides

1. Problem overview
   - What is evaluated (photo + annotations)
   - Assumptions and scoping
2. Method overview
   - Hypotheses, ...
   - How did you address threats to validity?
3. Pilot results
   - Photo + summary w bullets
4. Assessment
   - What works/doesn't in the study design?

Avoid ”feature creep”

You can do this



You can’t do this

Tips and challenges

- Use MyCourses on phone
- Use live notes as proxy measurements
- Pay attention to the task description. How do you operationalize “mobile” and “usability”?
- For quantitative measurements, how will you know if a particular value (e.g., 3.5.) is good/bad?
- Scope your study such that you can run a sensible pilot
- Minimal maximal pilot (MMP)
- Identify as many threats to validity you think are relevant. Then pick 2-3 most critical ones and focus on mitigating them