Empirical Software Engineering Research

Experiment Conduct

Norman Peitek, Annabelle Bergum, Lina Lampel, Sven Apel

Learning Goals

- Understand the phases of conducting empirical studies (tying in many topics from previous lectures)
- Being able to differentiate between a study and an experiment

Overview of Study Phases

1. Study planning, including experiment design are the starting point to learn the background knowledge, select study methods, plan the experiment, …
   - This often takes the longest amount of time
2. Selecting sample/recruiting participants
3. Study execution: collect data
4. Data analysis
5. Study report

Overview of Study Phases

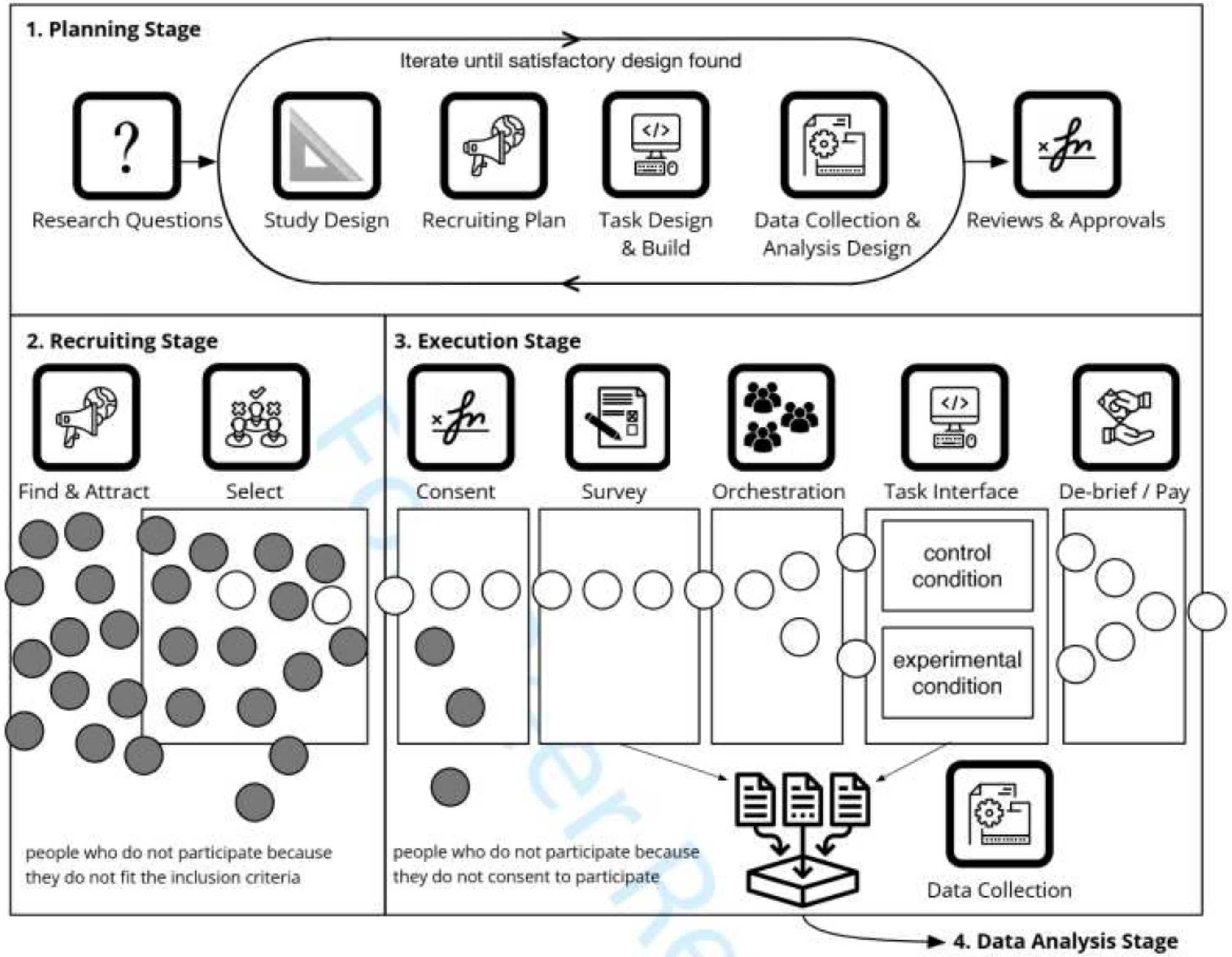


Fig. 1. Key stages in a programmer user study. Within each stage, activities may occur in various orders. This is an augmented version of Figure 1 of Ko et al. 2015 [54]

Study Planning

- So far, we discussed different designs on how to conduct controlled experiments
  - Qualitative studies will follow tomorrow
- Experiment design is a part of study design
  - There are other empirical research paradigms beyond experiments
- First priority: find the “right” research question
  - This can be the hardest part in empirical research
  - There is also little guidance on which research methods are suitable for what kind of problem (and how to choose)

Study Planning

If you have a research topic … what is the right research method?

- Often, convenience is the single most important factor
  - Method(s) that you are familiar with and can perform
  - Method(s) that are accepted in your field and have a chance to get published
  - Method(s) your supervisor/thesis advisor considers interesting/relevant
  - Method(s) for which you have tools/equipment/data available
- In addition, a close look at the research topic reveals some direction
  - "Why?" and "how?" questions regarding people often indicate qualitative methods
  - Robust causal explanations are ideal for experimentation
  - If there are lots of primary studies, secondary studies or replications can be of interest

Study Planning

Finding the "right" research question

- This is crucially important
  - If you ask the wrong question even an excellent answer does not help
- The research question and research method go hand in hand
  - This an iterative process over time
  - Often, we reduce from really interesting questions to (boring) simple questions that we can actually answer. Then, after some results we can ask slightly less simple questions
  - Some questions are very difficult/impossible to answer

Study Planning: PICOC-Approach

A rough heuristic is the PICOC approach

- Population: *who?*
- Intervention: *what/how?*
- Comparison: *compared to what*
- Outcome: *what are you trying to accomplish/improve?*
- Context: *in what kind of organization/circumstance?*

Do not take this as exact guide, this should only help you to get started.

Study Planning: Investigate Standards

- Do research before doing research!
- Often, comprehensive guides on how to conduct studies are available
  - Practical guide on eye tracking: https://link.springer.com/article/10.1007/s10664-020-09829-4
  - …
- A reverse perspective is looking at guidelines of evaluating research (ACM, review guidelines, …)
- Keep in mind there are no perfect studies and that deviations are always possible

Sample Selection

Selecting Sample from Population

- A population is the entire group that a researcher intents to draw conclusions on
- Often the entire population cannot be invited (unless it is very small, e.g., all CS profs at Uni Saarland), so researchers select an (representative) sample
- For example, before a federal election in Germany, Forsa, INSA and others sample a few thousand people (out of ~82 000 000) and can estimate the election results to a degree
    - They are very detailed about how they draw their sample of ~0.002%

Selecting Sample from Population

- Sampling is often necessary, because…
  - It is unfeasible to invite the entire population
  - It is more cost-effective to only invite a sample
  - It is practically easier to only invite a sample
  - The analysis method(s) are limited to smaller data sets
- A sampling error refers to the difference in characteristics between the population and the sample
  - A researcher invites a random sample of 100 people (mean age: 22 years, 80% female) to a study, which intends to draw conclusions on the population of Germany (mean age: 45 years, 50% female)

Selecting Sample Strategy

Participants can either

- be selecting randomly (often difficult, but better representation), or
- by convenience (easier data collection, but typically less representative)

Selecting Sample: Recruiting Human Participants

Imagine you are writing a thesis and must recruit participants for your small study.

What would be your strategy to find participants?

Selecting Sample: Recruiting Human Participants

- Recruiting strategy depends on the type of study
- Some possible strategies that are not mutually exclusive:
  - Trawling: pick up participants slowly over time
  - Fishing: pick up participants with incentives
  - Preying: pick up participants based on their location
  - Spear-fishing: pick up participants by specifically targeting suitable candidates
- Remember, participation of humans is entirely voluntary (will discuss on Friday in Ethics)

Selecting Sample: Recruiting Human Participants

Trawling: pick up participants slowly over time

- Is particular useful if you need many participants
- Typically, via bulletin boards, email lists, websites, …
- Teachers may announce studies in their classes
- Can easily be complemented with other strategies
- But, depending on the goals, trawling may not attract enough participants

Selecting Sample: Recruiting Human Participants

Fishing: pick up participants with incentives

- Is particular effective if you understand the needs of potential participants
- Pretty typical for research in universities
- Depending on the audience, it might be difficult to find incentives

Preying: pick up participants based on their location

- If you have a specific target group, random fishing via email lists etc. may not get you many responses → target a specific group
- For example, researchers studying Microsoft employees waited at their shuttle bus stop to hand out questionnaires that they filled out on their walk to the building

Selecting Sample: Recruiting Human Participants

Spear-fishing: pick up participants by specifically targeting suitable candidates

- Personal connection or introduction helpful
- If you do not have direct connections, find one who does
- Only possible if the n is very small, but can be effective

Snowballing: ask participants to find further participants

Selecting Sample: Recruiting Human Participants

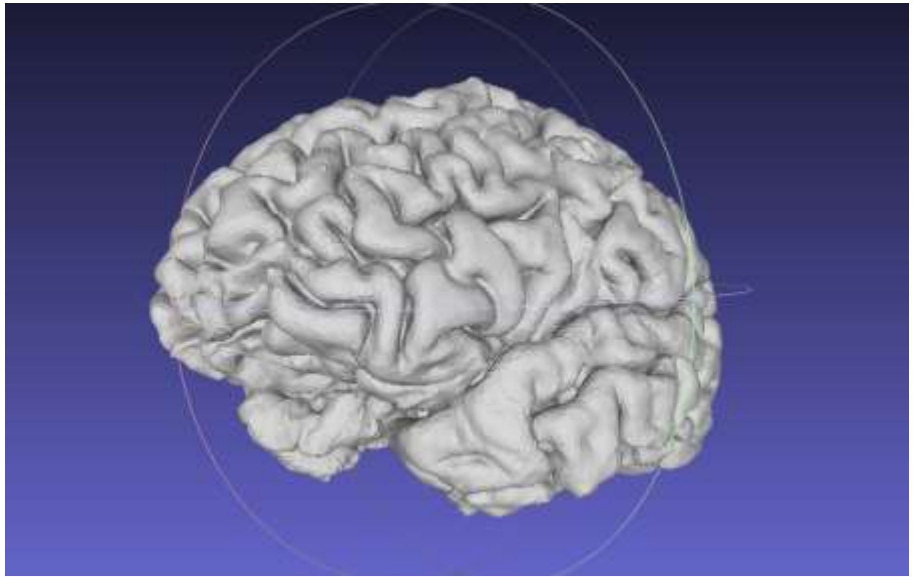
- Many studies feature students as participants
- Some peer reviewers immediately reject studies if the sample is only students
  - However, there is no generally valid reason to do so
  - Depending on the research question, students can be a representative or even exactly the right sample
  - Sometimes, a non-representative sample is also acceptable (e.g., for exploratory studies)
- Students are especially at risk of biasing the results
  - They may put extra effort into fulfilling expectations
  - They may have been primed by the researcher

Incentives for Human Participants

Have you participated in a research study? If yes, what motivated you to do so?

Incentives for Human Participants

- Money (if not too much) or extra credits for students are commonly used
  - Pre- versus post-payment
  - Small payment versus large lottery payment
- Interest participants how “cool” the study is and rely on good will
  - Cheap and effective, but hard to get right
- Little rewards: free chocolate, snacks, coffee, …
- Creative rewards: individual brain model



Experiment Execution: Task Instructions

- Every experiment requires to instruct the participants on what to do
  - Provide detailed information on how to do the experiment
  - Tip: test the instructions with friends/colleagues
- It is easy to unintentionally bias the participants → must be careful to avoid it
  - Oral explanations vary greatly, even if someone follows a script
  - Written or video instruction rather than personal explanation to reduce bias
- Interestingly, the task instruction as well as feedback can influence participants
  - For example, participants receiving a "WRONG!" might be less motivated than "Wrong answer, but you will get the next one ☺"

Sampling Strategies

	Procedure	Strength(s)	Weakness(es)
Convenience sampling	Select easiest ones	Least effort	May introduce bias
Random sampling	Select elements at random	Sample reliability relies only on sample size and population size	May result in unrepresentative samples
Systematic sampling	Select first element at random and then every nth element	Ensure representativity	Non-randomness may weaken confidence
Quota sampling	Divide population into sub-populations and sample those	Allows representative samples from inhomogeneous populations	Potentially large effort

Data Collection

Data Collection: Questionnaire

Imagine you are planning a study where participants need to fix a bug, either in Java or Python code. Afterwards, you want to collect their subjective view through five questions.

Do you ask the five questions through an online form or personally ask them (and write them down)?

What are advantages and disadvantages of each approach?

Data Collection: Audio

Audio recording of interviews is easy nowadays

- We still have tape recorders at LIN
- Nowadays, a phone is sufficient (or zoom recording, if online)

Transcription is expensive and can pose privacy concerns

- Automated transcription systems are getting better, but struggle with two/multiple speakers

Data Collection: Sensors

Sensor devices for physiological or neuroimaging automatically “enter” data.

- However, researchers must still ensure the data is saved correctly
  - Particularly important are time stamps, especially with multi-modality
  - For example, if one wants to connect eye-tracking with EEG data, one must be able to align them on the same timeline
  - Their internal timer might be different
- Data cleaning/outlier removal is necessary

Data Analysis

- For details on data analysis: see previous lectures
- (Unintentional) p-hacking
  - Wrong: Running many statistical tests and focusing/only report on the statistically significant ones
  - Right: Correct for multiple testing and report the complete analysis
- HARKing (hypothesizing after the results are known)
- Double-dipping
  - Use the same data set for exploration and confirmation

Reporting on Experiment

- Will discuss in detail on Friday in “Report Writing”

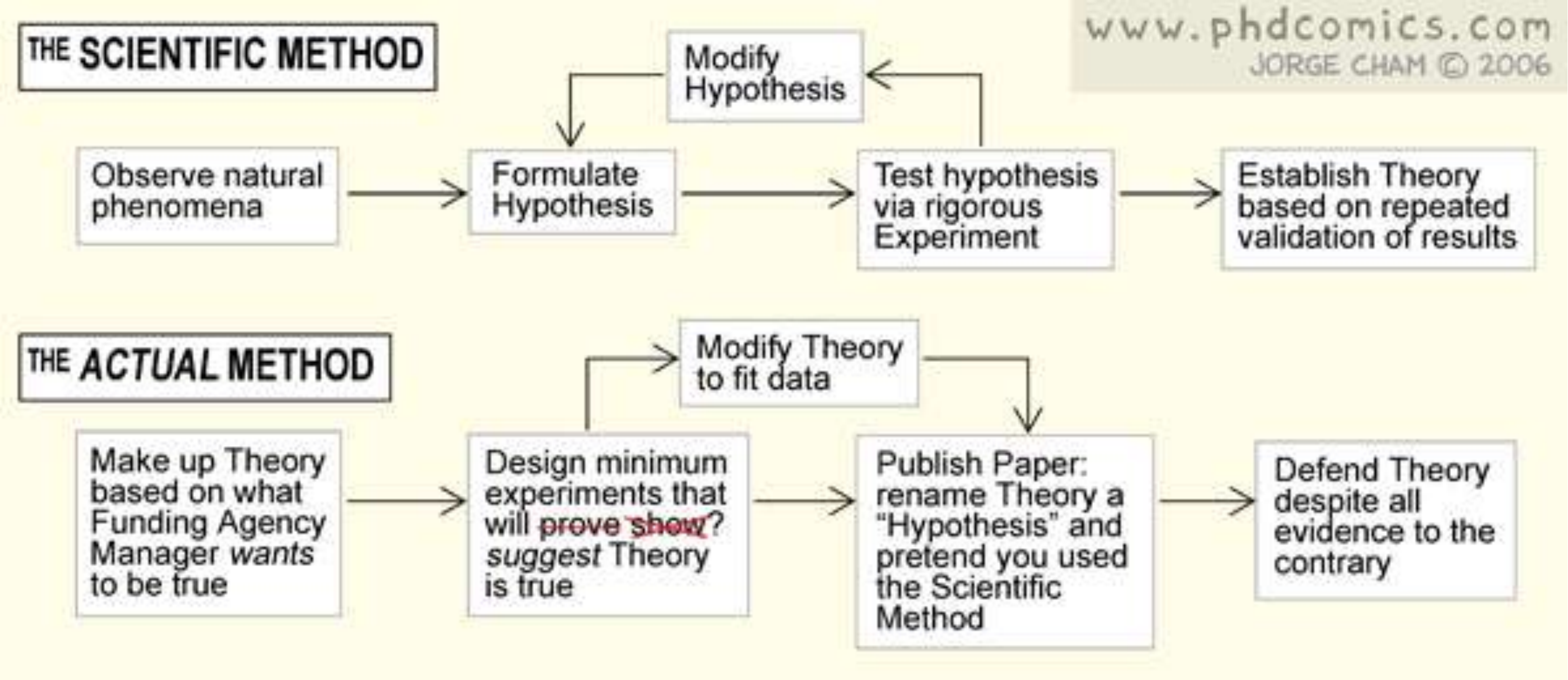
Empirical Software Engineering Research

How to Mess Up Your Experiment

Norman Peitek, Annabelle Bergum, Lina Lampel, Sven Apel

Learning Goals

- Some people learn more from anti patterns than patterns, so let's discuss some anti patterns of experiments ☺


www.phdcomics.com
JORGE CHAM © 2006
THE SCIENTIFIC METHOD
Modify Hypothesis
Observe natural phenomena
Formulate Hypothesis
Test hypothesis via rigorous Experiment
Establish Theory based on repeated validation of results
THE ACTUAL METHOD
Modify Theory to fit data
Make up Theory based on what Funding Agency Manager *wants* to be true
Design minimum experiments that will ~~prove~~ ~~show~~? *suggest* Theory is true
Publish Paper: rename Theory a "Hypothesis" and pretend you used the Scientific Method
Defend Theory despite all evidence to the contrary

Study Objectives

- No formulation of objectives (through research questions/hypotheses)
  - Select completely useless goals
  - Fail to iteratively improve research objectives
- Fail to specify and operationalize variables
- Ignore (boring) advice from your advisors
  - They do not intend to crush your ideas; they want to help you succeed

Study Design

- Include too many variables (e.g., five (in)dependent variables)
- Select dependent variable that is not influenced by independent variable at all
- Fail to consider important confounding factors
- Unsuitable experiment design (e.g., within-subject design to compare groups)
- Ignoring identified problems during the design phase („will deal with them later on“)
  → does not work in empirical research

Study Conduct

- Skip pilot testing
- Skip technical testing
- No plan for data collection (e.g., calendar, resource management, …)
- No backups/safe storage of data

Data Collection

- Collected data (method) does not fit the research question
- Different task introduction to each participant
- Unsuitable participant recruiting
- Too few samples/participants
- Participants unwilling/unable to understand/follow instructions
- Continuing the experiment when things go wrong

Data Analysis

- Insufficient data cleaning, ignoring outliers
- Throwing out data points in a biased way (“to fit the story”?)
- Make data analysis messy (read: use Excel)
- Backup scripts, …
- Inappropriate/misleading descriptive statistics
  - Or, no reporting of descriptive statistics at all

Data Analysis

- Applies incorrect inferential statistics
- Failing to visualize data
- Just test for statistical significance, ignoring effect sizes
- Claim causality while only testing for correlation
- Misleading visualizations (talk more about it on Friday)
- p-hacking, HARKing

Reporting

- Unclear language due to writing clarity and missing terms
- Lack of detail in methodology
- Fail to report descriptive and inferential statistics
- Overclaim results („invited one student who hates experiment → outlaw experiments forever“)
- Start writing the night before the deadline
- Fail to proof-read and edit your manuscript
    - Ideally, have some time in between being “done” and submission

Final Note

I... HAVE A CONFESSION TO MAKE, PROF. SMITH.
UH OH.
I HAD AN ACCIDENT IN THE LAB.
DID ANYONE DIE?
NO.
IS ANYONE GOING TO SUE US?
NO.
WAS ANY EQUIPMENT PERMANENTLY DAMAGED?
NO.
DID YOU CLEAN IT UP?
OF COURSE!
JORGE CHAM © 2013
DID YOU GET DATA OUT OF IT?
ACTUALLY, YES.
CAN YOU DO IT AGAIN?
WWW.PHDCOMICS.COM

Further Reading

- https://raw.githubusercontent.com/SIGPLAN/empirical-evaluation/master/checklist/checklist.pdf
- https://elemental.medium.com/when-science-needs-self-correcting-a130eacb4235