



UiO : **Department of informatics**
University of Oslo

Menneske-maskin: Kvantitative metoder

Dag Sjøberg and Gunnar Bergersen



About Me



- Current position: Professor at University of Oslo
 - Software Process Improvement, Agile and Lean Methods, Software Quality, Empirical Research Methods
- Education:
 - MSc, University of Oslo
 - PhD, University of Glasgow, Computing Science
- Prior work experience
 - National University Hospital (Rikshospitalet)
 - SINTEF ICT
 - Simula Research Laboratory
 - Statistics Norway (SSB)
- Startup
 - Member of steering committee and co-owner of four startup companies

About me



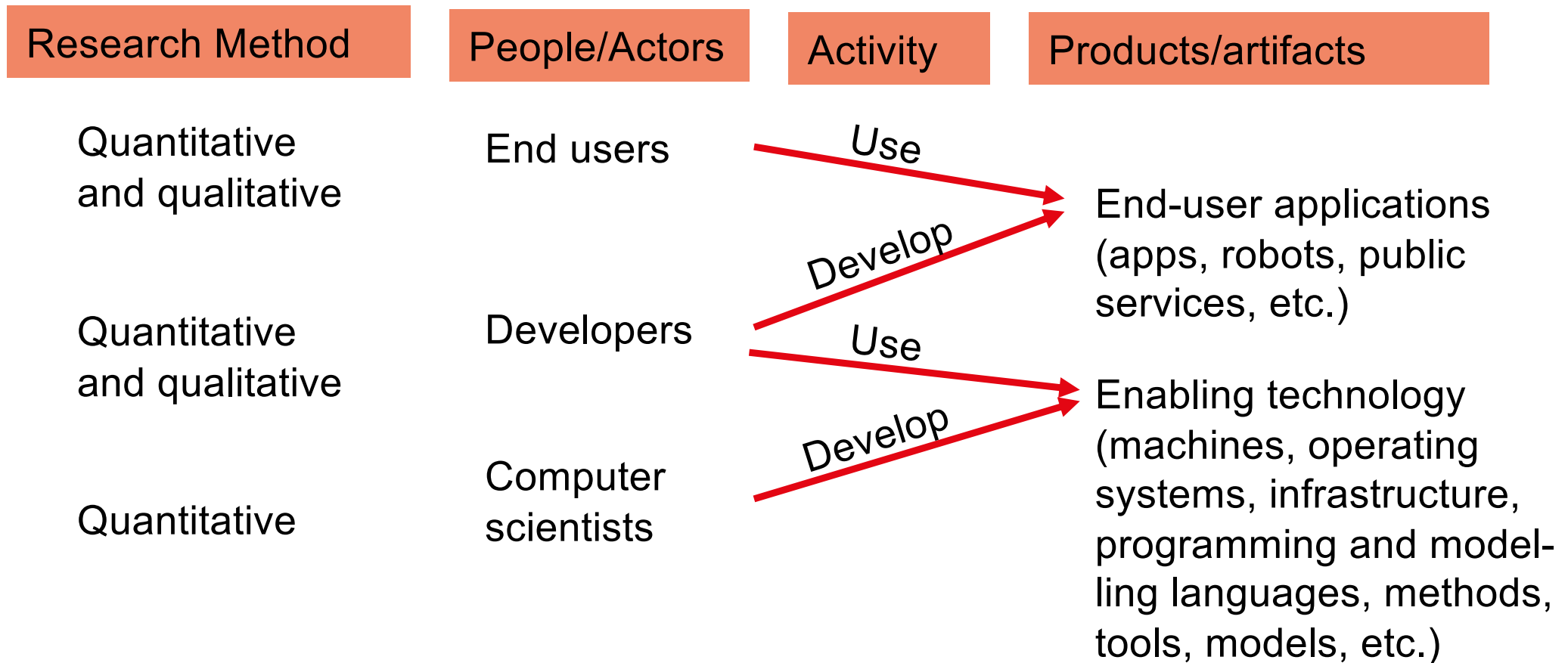
- Current:
 - Associate professor in Software Engineering
- Education
 - MSc (2001) and PhD (2015) from University of Oslo
 - PhD thesis: “Measuring programming skill”
- Prior work experience
 - Programmer
 - IT Project leader of two companies
 - CEO of three companies and Chief Product Officer in Greps (skill testing of developers)

Research method = the way knowledge is created

Good method → **Reliable knowledge**

Poor method → **Unreliable knowledge**
(most extreme: fake news, lies)

Typical pattern of informatics/computer science



Writing a Master's thesis

You may wish to

- propose or develop a new X (process, method, technique, practice, language, tool, framework, algorithm, robot, etc.) that is supposed to be better than what exists,
- find out whether an existing X is better than an existing Y, or
- how to improve X

Most Master's theses would include some of this

- Thesis: A scientific statement
- Master (or PhD thesis): A justification of that statement

What does it mean to be better?

- How much?
- How many?
- How large?
- How old?
- How long?
- How high?
- How warm?
- How thick, firm, etc.

Answers are measured in terms of *quantitative data*

Quantitative data

- Data expresses quantity
- Data expressed as numbers
- Used in statistics

Mentimeter:
What kind of
data do you
plan to collect
for your thesis?



Qualitative data

- Data expresses quality in some sense
- Data expressed as text, images and forms except numbers
- Can obtain quantitative data indirectly if a mapping exists from quantitative to quality data
- Not used in statistics

Quantitative empirical methods

- Experiments and surveys typically collect quantitative data
- Therefore called “quantitative empirical methods”

Empirical means using evidence based on observation or experience rather than theory or pure logic

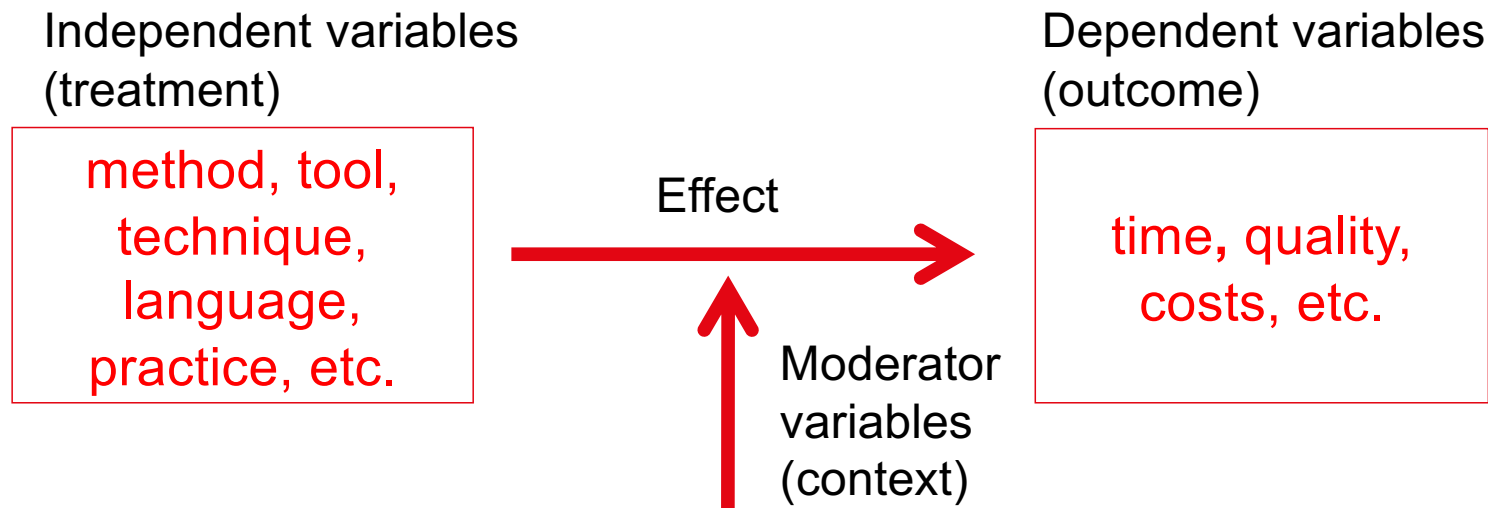
How to find out whether something is better?

Investigate what works best in practice, that is, perform an empirical study

- ❖ Measurements/Statistics
- ❖ Experiment
- ❖ Survey: people are asked about their opinions
- ❖ Other studies:
 - case studies, possibly using interviews
 - action research
 - ethnographic studies
 - others

**What do you plan to
investigate in your thesis?**

(Controlled) Experiment



- An experiment is a cause-effect study, that is, an intervention (treatment) is introduced to observe its effects
- Difference in the outcome is supposedly caused by the different treatments (or treatment compared with no treatment, that is, the control group)

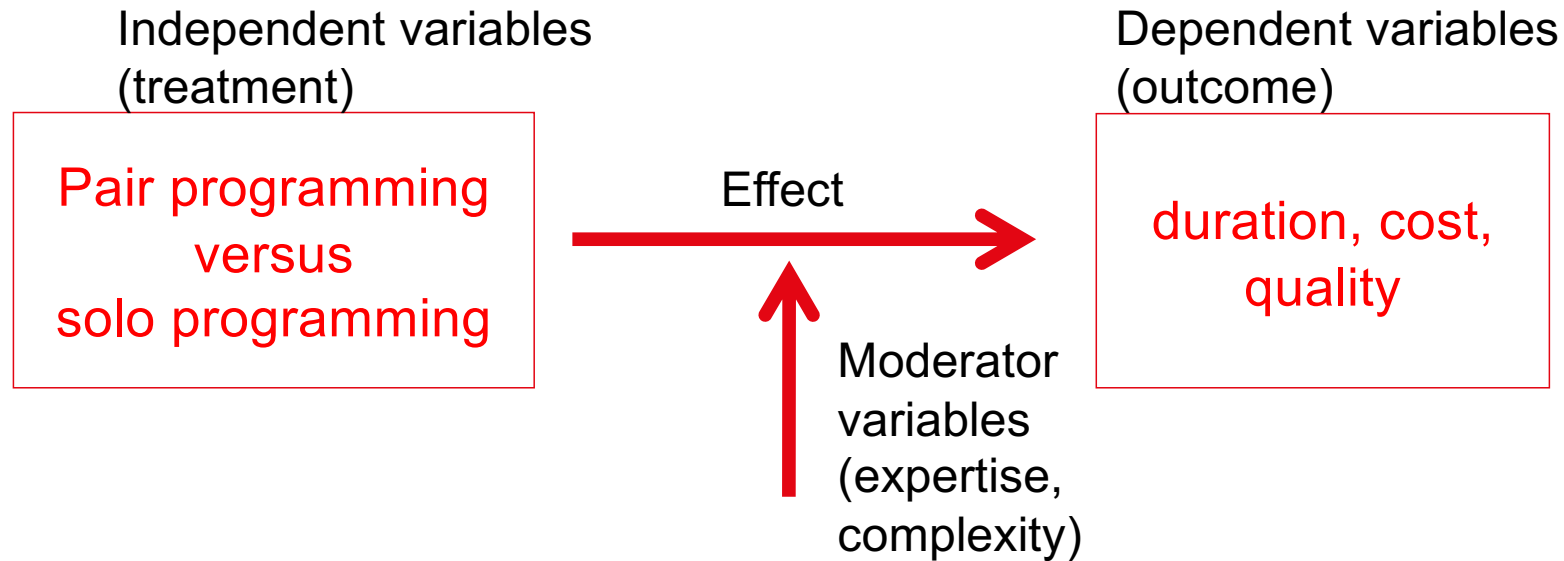
What is best?

Pair programming or solo programming

- 295 junior, intermediate and senior **professional Java consultants** from 29 companies were paid to participate (one work-day)
- 99 individuals; 98 pairs
- The pairs and individuals performed the same Java maintenance tasks on either:
 - a "simple" system (centralized control style), or
 - a "complex" system (delegated control style)
- We measured:
 - duration (elapsed time)
 - effort (cost)
 - quality (correctness) of their solutions



Experiment



Results

Programmer Expertise	Task Complexity	Use PP?	Comments
Junior	Easy	Yes	Provided that increased quality is the main goal
	Complex	Yes	Provided that increased quality is the main goal
Intermediate	Easy	No	
	Complex	Yes	Provided that increased quality is the main goal
Expert	Easy	No	
	Complex	No	Unless you are sure that the task is too complex to be solved satisfactorily even by solo seniors

**One cannot state that pair programming
is generally beneficial or not!**

Other examples of experiment

- Two algorithms may be compared regarding, for example:
 - Their ability to solve a task
 - Their performance (speed)
 - Their energy consumption
 - Their understandability

A/B testing



- Background
 - Historically used in marketing, design, games
 - Data-driven product development & profitability tuning
 - Assumption: "We're all wrong most of the time"
- Typical use
 - Fast release cycles
 - Bottom-up and context dependent (for example, "I want to improve X for part Y")

Requires

- Two (or more) version to be compared (A, B, ...)
- At least one outcome variable (KPI or other metric or measure) to improve

Experiments

Advantages

- They are a well-established strategy, seen by many as the most 'scientific' approach
- The only research strategy that can prove cause-effect relationships
- Laboratory experiments permit high levels of precision in measuring outcomes and in analyzing data



Disadvantages

- Laboratory experiments often create artificial situations that are not comparable to real-world situations
- Often difficult or impossible to control all the relevant variables
- It is often difficult to recruit a representative sample of participants



Surveys



Common in society

- Requires relatively **few resources** to include many people
- Create statistics and test hypotheses over characteristics of the target group (the population being investigated)
- Obtain information about people's **opinion** about what, how much, how many, how and why or what people say they do
- As opposed to case studies and ethnography, one does *not observe*

Advantages

- They provide a wide and inclusive coverage of people or events
- They can be administered from remote locations using mail, email or telephone
- They can provide a lot of data in a short time at a reasonable cost
- They can be quantitatively analysed
- They can be replicated



Disadvantages

- They lack depth
- They tend to focus on what can be counted or measured
- They do not establish cause-effect
- They cannot judge the accuracy or honesty of people's responses, for example, by observing their body language



Mentimeter: What kind of research method do you plan to use for your thesis?



Hypothesis testing

Null hypothesis:

- "there is no difference between the effect of treatments (experiments) or between groups (surveys)"
- "there is no difference between solo and pair programming"

P-value

- If it's very unlikely, for example $< 5\%$ (significance level), that we had obtained the results that we actually got, if the null hypothesis were true, then we reject the null hypothesis and claim the alternative hypothesis ("there *is* a difference ...")
- The likelihood that we obtained the results we did assuming the null hypothesis is true, is called the p-value (probability value)
- One uses statistical methods to test null hypotheses

How likely is 5 %?

- Approximately the probability of exactly 8 heads in 10 tosses is
 $P(X=8) = (10 \text{ choose } 8)/2^{10} = 45/1024 \approx 0.044$

Effect size

- P-values is about how likely it is that there is a difference
- Effect size is about how large the difference is
 - Is the difference large enough to have any meaning in practice?

V.B. Kampenes, T. Dybå, J.E. Hannay and D.I.K. Sjøberg. A Systematic Review of Effect Size in Software Engineering Experiments, *Information and Software Technology* 49(11-12):1073-1086, 2007

Validity of empirical studies

- **Internal validity**

- Is the difference that we observed between the groups that received the treatments actually caused by the treatments, or may there be other causes for the difference?

- **Construct validity**

- If a complex concept is measured, does the measurements represent the concept in a satisfactory way? For example, is it OK to measure quality of a software system only in terms of number of bugs?

- **External validity**

- Can we generalize the results we found; that is, is it likely that we would obtain the same results in other settings?

- **Statistical conclusion validity**

- Is correct statistics used?

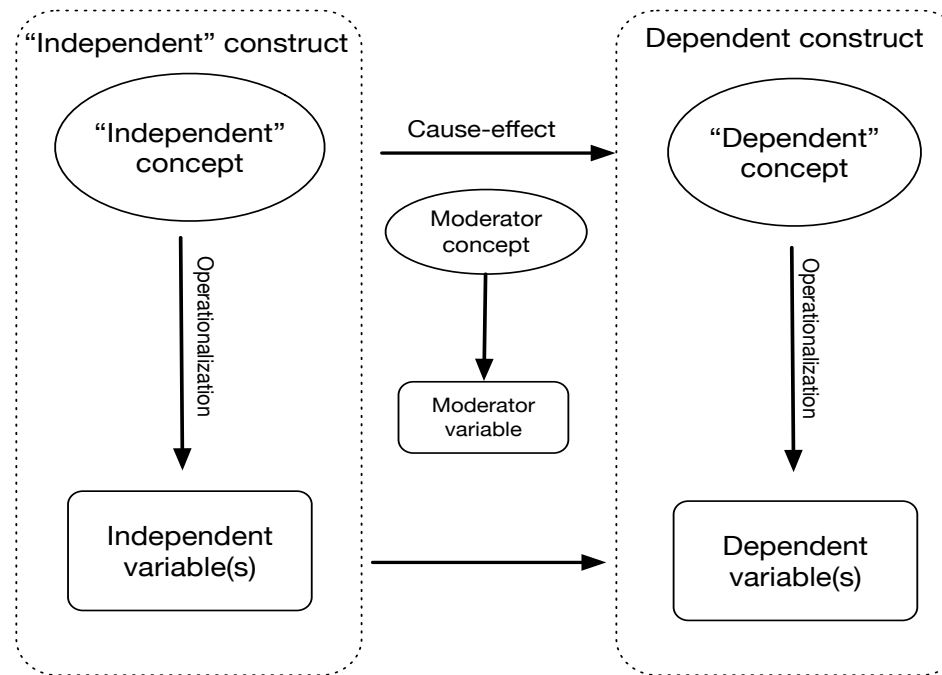
Construct validity →

Measurements

**Conceptual
level**

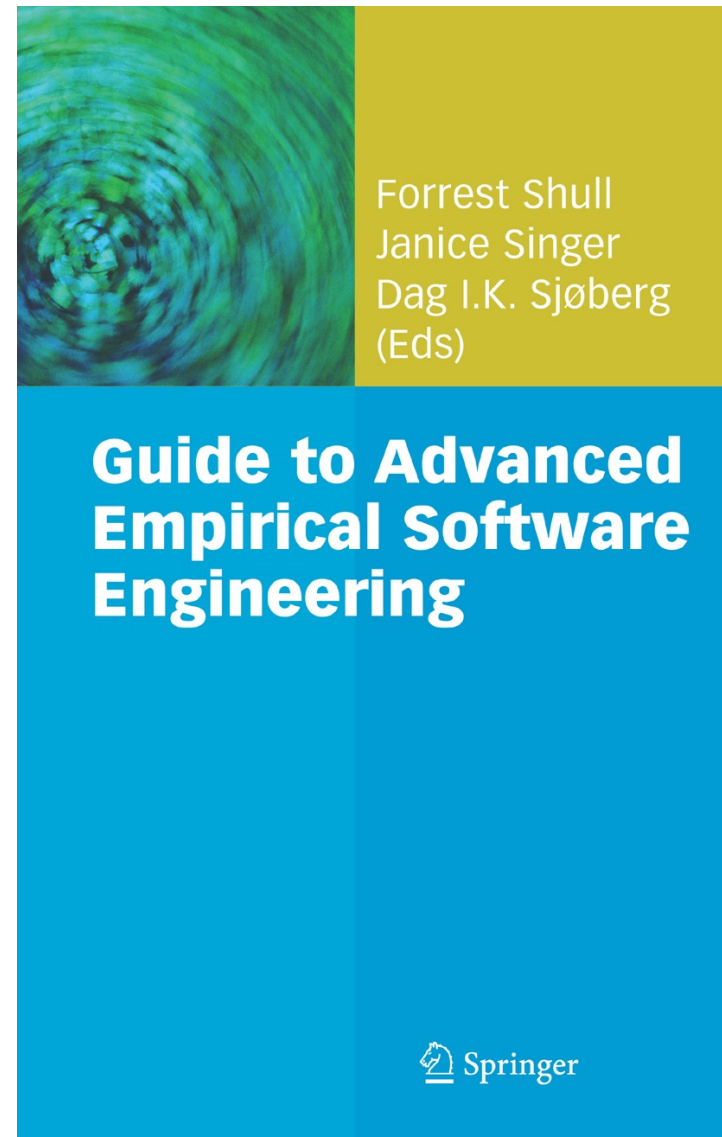
**Operational
level**

Internal validity →



Much literature on research methods

Dag Sjøberg and Gunnar Bergersen





UiO : **Department of informatics**
University of Oslo

Good luck with your thesis work!

