Second intro week Master's, 12 January 2024

Quantitative empirical methods

Dag Sjøberg and Gunnar Bergesen

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)

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 should include some of this

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

How to find out whether something is better?

- Investigate what works best in practice, that is, perform an empirical study
  - Experiment
  - Survey: people are asked about their opinions
  - Other studies:
    - case studies, possibly using interviews
    - action research
    - ethnographic studies
    - others

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

Structure

- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

What does it means 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

In your MSc thesis, you may wish to

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

- **What do you want to investigate in your thesis?**

Structure

- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

You may wish to

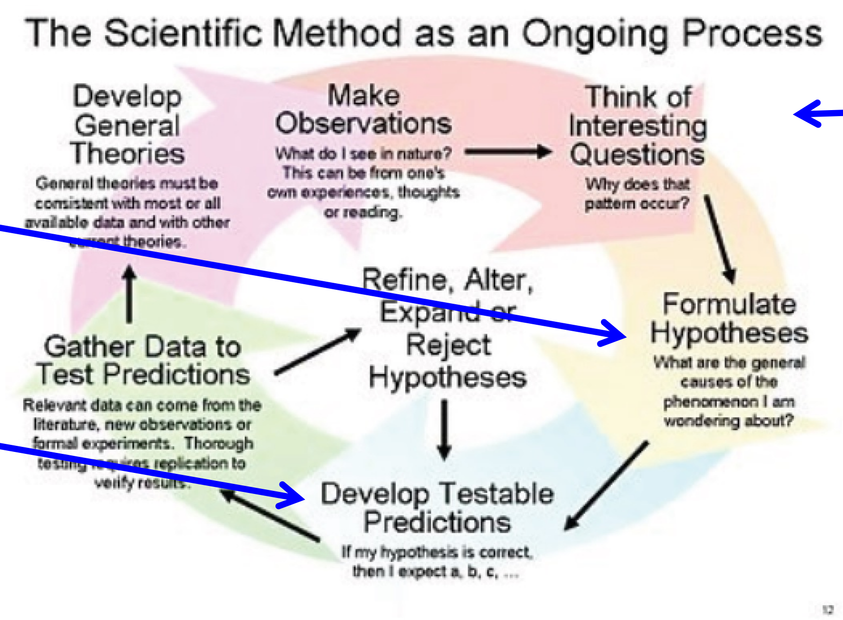
- propose or develop a new method, tool, technique, language, practice, etc. that is supposed to be better than what exists, or
- find out whether an existing X is better than an existing Y, or
- how to improve X

What do you believe is the case?
That is, formulate a hypothesis

Test the hypothesis: given that the hypothesis is correct, which results do you expect (predict)?

If the hypothesis is confirmed, what you believed is not hypothetical any more; you have a **thesis**

Scientific method: Hypotheses and testing (or conjectures and refutations)

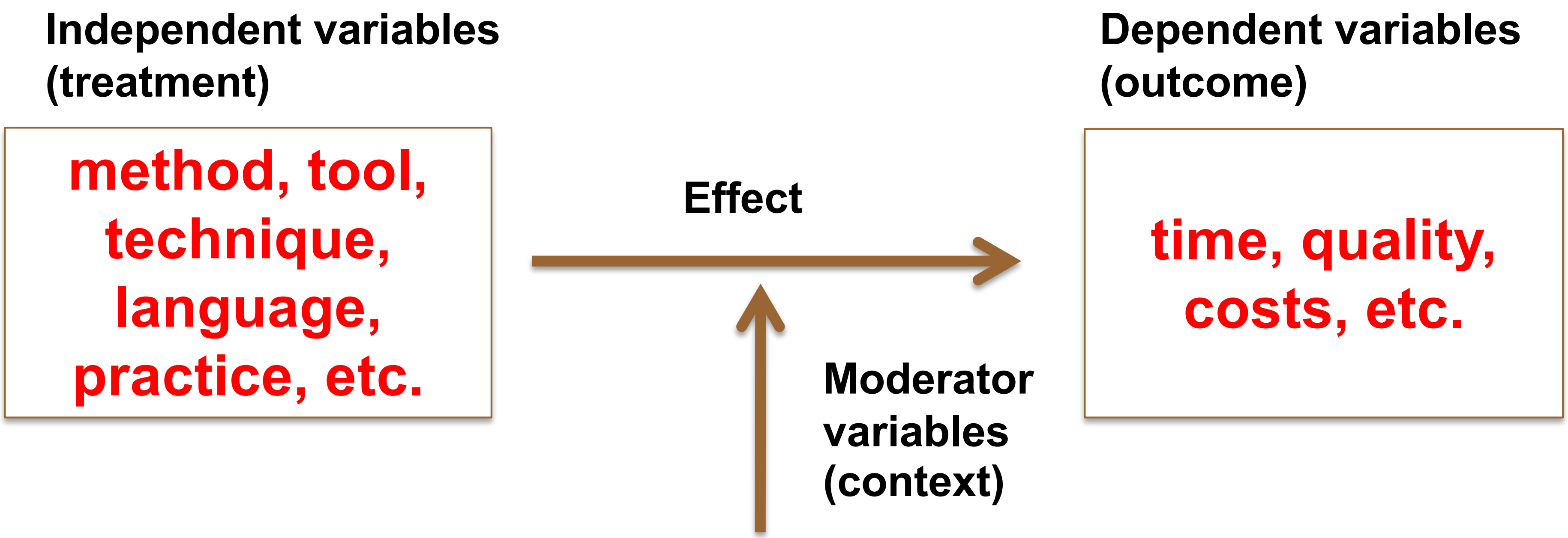


Bygstad 2018

Structure

- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

Experiment



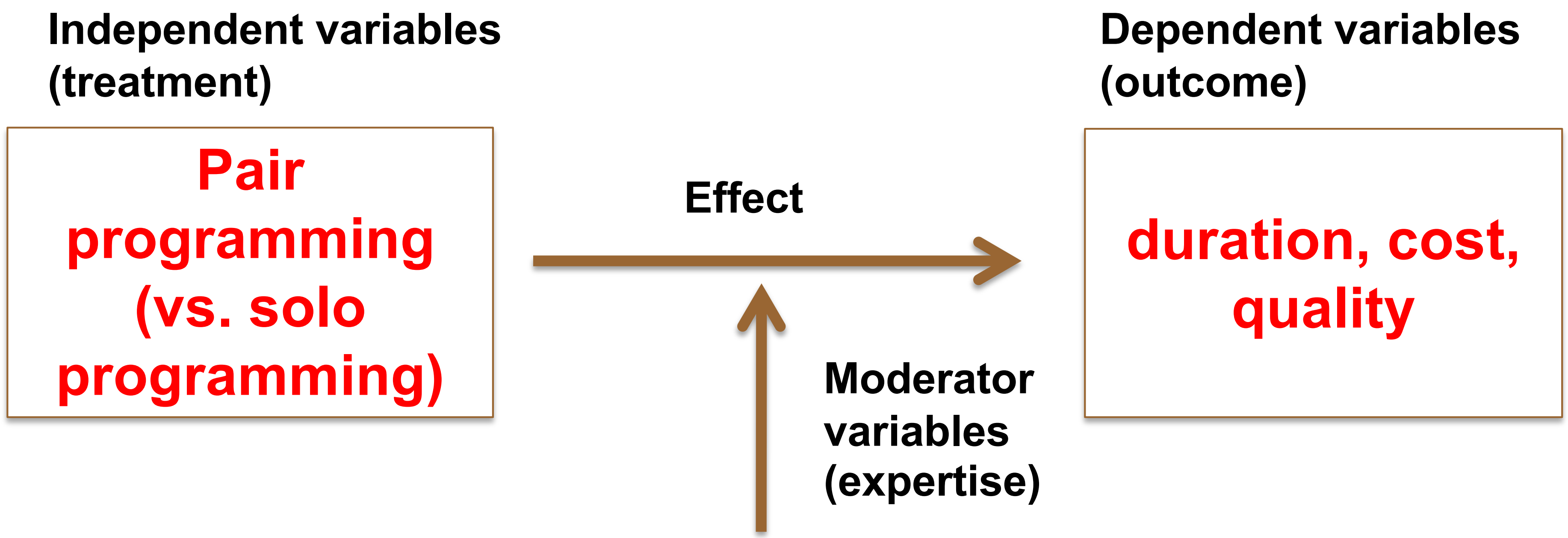
- An experiment is a cause-effect study, that is, an intervention (treatment) is introduced to observe its effects
- Difference in the outcome is supposed to be caused by the different treatments (or by the treatment compared to no treatment, that is, the control group)

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

*E. Arisholm, H. Gallis, T. Dybå, and D. Sjøberg, “Evaluating Pair Programming with Respect to System Complexity and Programmer Expertise,” *IEEE Transactions on Software Engineering*, 2007, 33(2): 65-86.

Experiment


Independent variables (treatment)
Pair programming (vs. solo programming)
Effect
Dependent variables (outcome)
duration, cost, quality
Moderator variables (expertise)

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

The question of whether PP is beneficial or not in general, is meaningless!

Other examples of experiment

- Two robots may be compared regarding, for example:
  - The time they need to solve a task
  - The speed, stability, elasticity, etc., of their movement
- Two algorithms may be compared regarding, for example:
  - How good they are at solving a task
  - Their performance (speed)
  - Their energy consumption
  - Their understandability

Structure

- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

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 (i.e., ”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

Structure

- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

Surveys


Extremely Satisfied
2. To what extent do you agree with the
Disagree Completely
Strongly Disagree
Somewhat Disagree
Somewhat Agree
Strongly Agree
Agree Completely
working for

Common in society

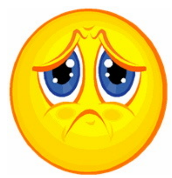
- Requires relatively few resources to include many people
- Create statistics and test hypotheses over characteristics of the target group (the population being investigated)
- Obtain information about people's *opinion* about what, how much, how many, how and why or what people *say* they do
  - As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

Advantages

- They provide a wide an 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 by observing their body language



Structure

- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

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 assumring the null hypothesis is true, is called the p-value (probability value)
- One uses statistical methods to test null hypotheses

Effect size

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

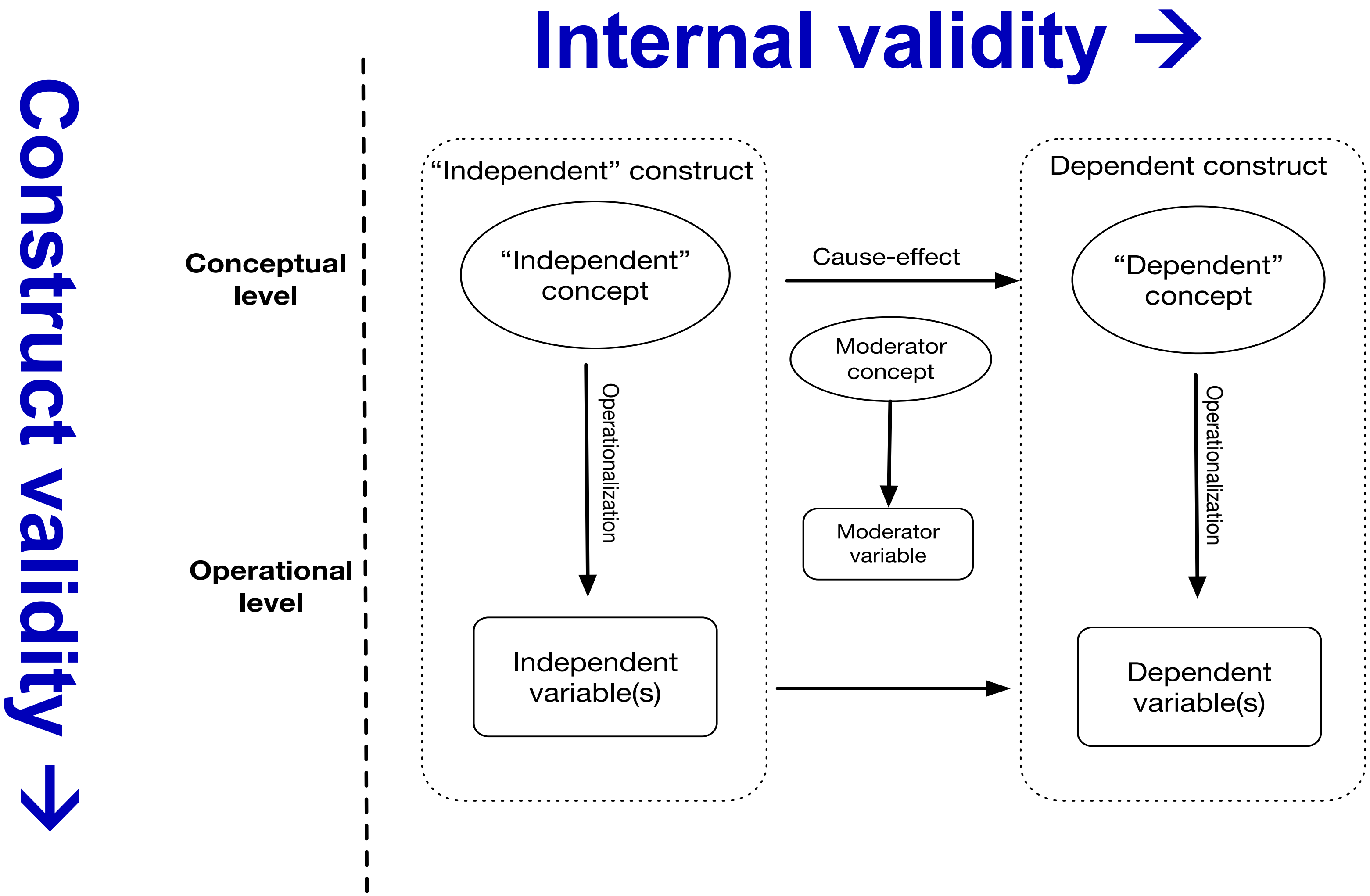
V.B. Kampenes, T. Dybå, J.E. Hannay and D.I.K. Sjøberg. A Systematic Review of Effect Size in Software Engineering Experiments, *Information and Software Technology* 49(11-12):1073-1086, 2007

Structure

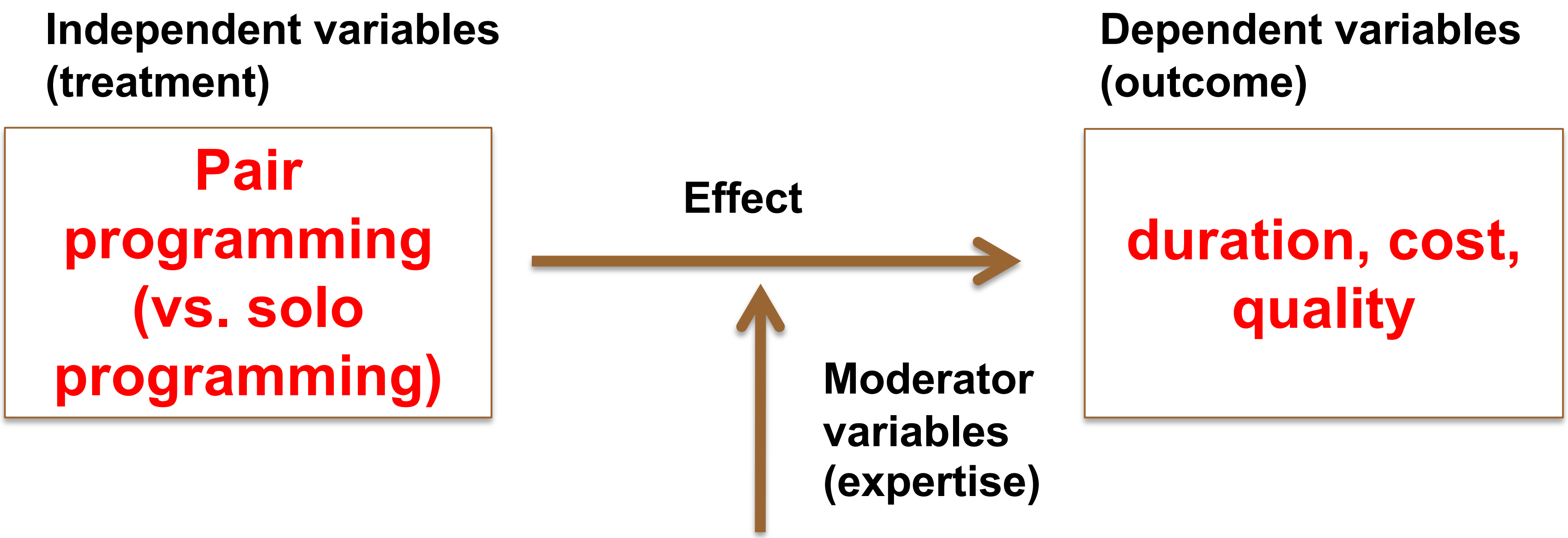
- Quantitative vs. qualitative data
- The research life cycle
- Controlled experiments
  - Experiment example
  - AB-Experiments
- Surveys
- Quality of experiments and surveys
  - Hypothesis testing and effect size
  - Validity

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 measure 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 found?
- **External validity**
  – Can we generalize the results we found; that is, is it likely that we would obtain the same result in other settings?
- **Statistical conclusion validity**
  – Is correct statistics used?


Internal validity →
Construct validity ↓
Conceptual level
Operational level
"Independent" construct
"Independent" concept
Operationalization
Independent variable(s)
Cause-effect
Moderator concept
Moderator variable
Dependent construct
"Dependent" concept
Operationalization
Dependent variable(s)

Experiment



Cannot generalize to different levels of expertise

External validity – to which population can we generalize?

- Sampling – Representativeness
  - https://www.aftenposten.no/viten/i/OnK87b/Psykologiforskning-gjelder-bare-for-noen-fa-mennesker--Nina-Kristiansen
  - "Nitti prosent av deltagerne i psykologistudier kommer fra Europa og USA, men de utgjør bare 18 prosent av verdens befolkning, … forskere i psykologi dessuten henter inn studenter som deltagere i studiene sine. Disse hvite, urbane ungdommene er ikke engang representative for befolkningen i sitt eget land, mener Gurven."

Empirical research methods – literature


The Future of Empirical Methods in Software Engineering Research

Dag I. K. Sjøberg, Tore Dybå and Magne Jørgensen

Dag I.K. Sjøberg received the MSc degree in computer science from the University of Oslo in 1987 and the PhD degree in computing science from the University of Glasgow in 1993. He has five years of industry experience as a consultant and group leader. He is now research director of the Department of Software Engineering, Simula Research Laboratory, and a professor of software engineering in the Department of Informatics, University of Oslo. Among his research interests are research methods in empirical software engineering, software processes, software process improvement, software effort estimation, and object-oriented analysis and design. He is a member of the International Software Engineering Research Network, the IEEE, and the editorial board of Empirical Software Engineering.

Tore Dybå received the MSc degree in electrical engineering and computer science from the Norwegian Institute of Technology in 1986 and the PhD degree in computer and information science from the Norwegian University of Science and Technology in 2001. He is the chief scientist at SINTEF ICT and a visiting scientist at the Simula Research Laboratory. Dr. Dybå worked as a consultant for eight years in Norway and Saudi Arabia before he joined SINTEF in 1994. His research interests include empirical and evidence-based software engineering, software process improvement, and organizational learning. He is on the editorial board of Empirical Software Engineering and he is a member of the IEEE and the IEEE Computer Society.

Magne Jørgensen received the Diplom Ingeneur degree in Wirtschaftswissenschaften from the University of Karlsruhe, Germany, in 1988 and the Dr. Scient. degree in informatics from the University of Oslo, Norway in 1994. He has about 10 years industry experience as software developer, project leader and manager. He is now professor in software engineering at University of Oslo and member of the software engineering research group of Simula Research Laboratory in Oslo, Norway. His research focus is on software cost estimation.

Future of Software Engineering(FOSE'07)
0-7695-2829-5/07 $20.00 © 2007 IEEE

358

IEEE COMPUTER SOCIETY


Empir Software Eng (2011) 16:425–429
DOI 10.1007/s10664-011-9163-y

Qualitative research in software engineering

Tore Dybå · Rafael Prikladnicki · Kari Rönkkö · Carolyn Seaman · Jonathan Sillito

Published online: 28 May 2011
© Springer Science+Business Media, LLC 2011

Qualitative research methods were developed in the social sciences to enable researchers to study social and cultural phenomena and are designed to help researchers understand people and the social and cultural contexts within which they live (Denzin and Lincoln 2011). The goal of understanding a phenomenon from the point of view of the participants and its particular social and institutional context is largely lost when textual data are quantified. Taylor and Bogdan (1984) point out that qualitative research methods were designed mostly by educational researchers and other social scientists to study the complexities of human behavior (e.g., motivation, communication, difficulties in understanding). According to these authors, human behavior is clearly a phenomenon that, due to its complexity, requires qualitative methods to be fully understood, since much of human behavior cannot be adequately described and explained through statistics and other quantitative methods. Examples of qualitative methods are action research, case study research, ethnography, and grounded theory. Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions and reactions.

Many in the software industry recognize that software development also presents a number of unique management and organizational issues that need to be addressed and solved in order for the field to progress. And this situation has led to studies related not only

T. Dybå
SINTEF, Trondheim, Norway

R. Prikladnicki
PUCRS, Porto Alegre, Brazil

K. Rönkkö
Blekinge Institute of Technology, Karlskrona, Sweden

C. Seaman (✉)
University of Maryland Baltimore County, Baltimore, MD, USA
e-mail: cseaman@umbc.edu

J. Sillito
University of Calgary, Calgary, Alberta, Canada

Springer