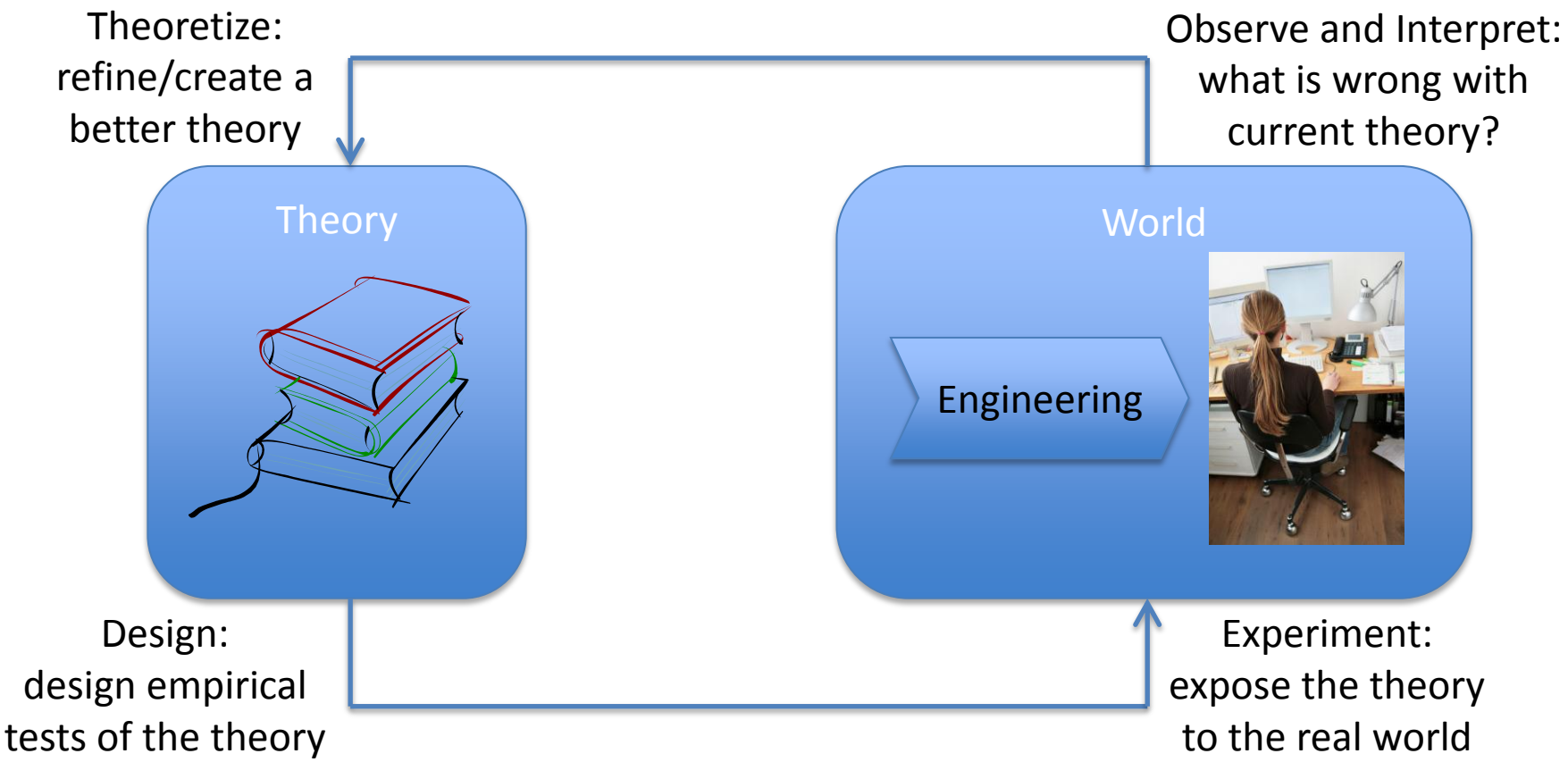
Validity Threats

Samuel Fricker | sfr@bth.se | 111130

www.bth.se

BLEKINGE INSTITUTE OF TECHNOLOGY

Research


Theoretize:
refine/create a
better theory
Observe and Interpret:
what is wrong with
current theory?
Theory
World
Engineering
Design:
design empirical
tests of the theory
Experiment:
expose the theory
to the real world

Every Research is Flawed

- What are potential problems?
- How can we plan research and report results to
  - be ethically correct?
  - remain credible?

Contents

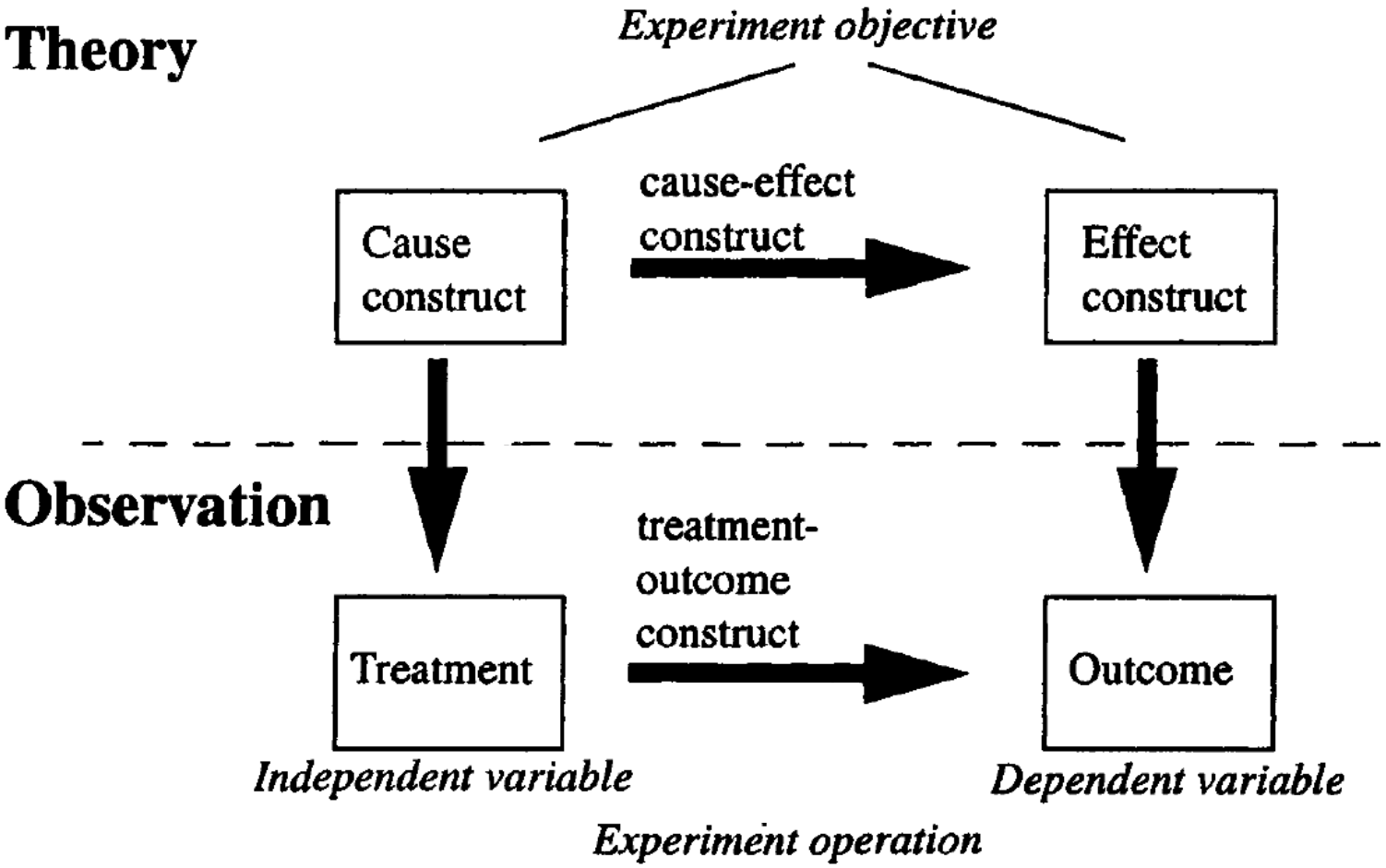
Understand

- Threats to validity

Agenda

- Threats to Validity in Experimentation
- Threats to Validity in Other Research Methods

Theory vs. Observation



Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Example: Requirements Prioritization Study

- Research Questions:
  Which prioritization technique is superior?
  - Pair-wise comparisons
  - Planning game
- Constructs: superior =
  - Time to conclude prioritization is shorter
  - Technique is easier to use
  - Prioritization results are more accurate

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Example: Requirements Prioritization Study

- Pairwise Comparisons (PWC): Analysis

Weight	Count	ID	Item	Preferences				
30%	3	a	Acquisition	a	d	a	a	
20%	2	b	Debt Collection	e	b	b		
20%	2	c	VAT	c	c			
10%	1	d	Reporting	e				
20%	2	e	Interfaces					

Pairwise Comparisons

Saaty (1980): The Analytical Hierarchy Process. McGraw-Hill. Böhm, Fuchs, Pacher (1993): System-Entwicklung in der Wirtschaftsinformatik. v/d/f.

Example: Requirements Prioritization Study

- Planning Game (PG):
  Dialogue between customer and developers

1. Customer lists requirements.
2. Developers estimate risk and effort, while customer clarifies.
3. Customer selects the highest-value set of requirements that are feasible withing given budget.

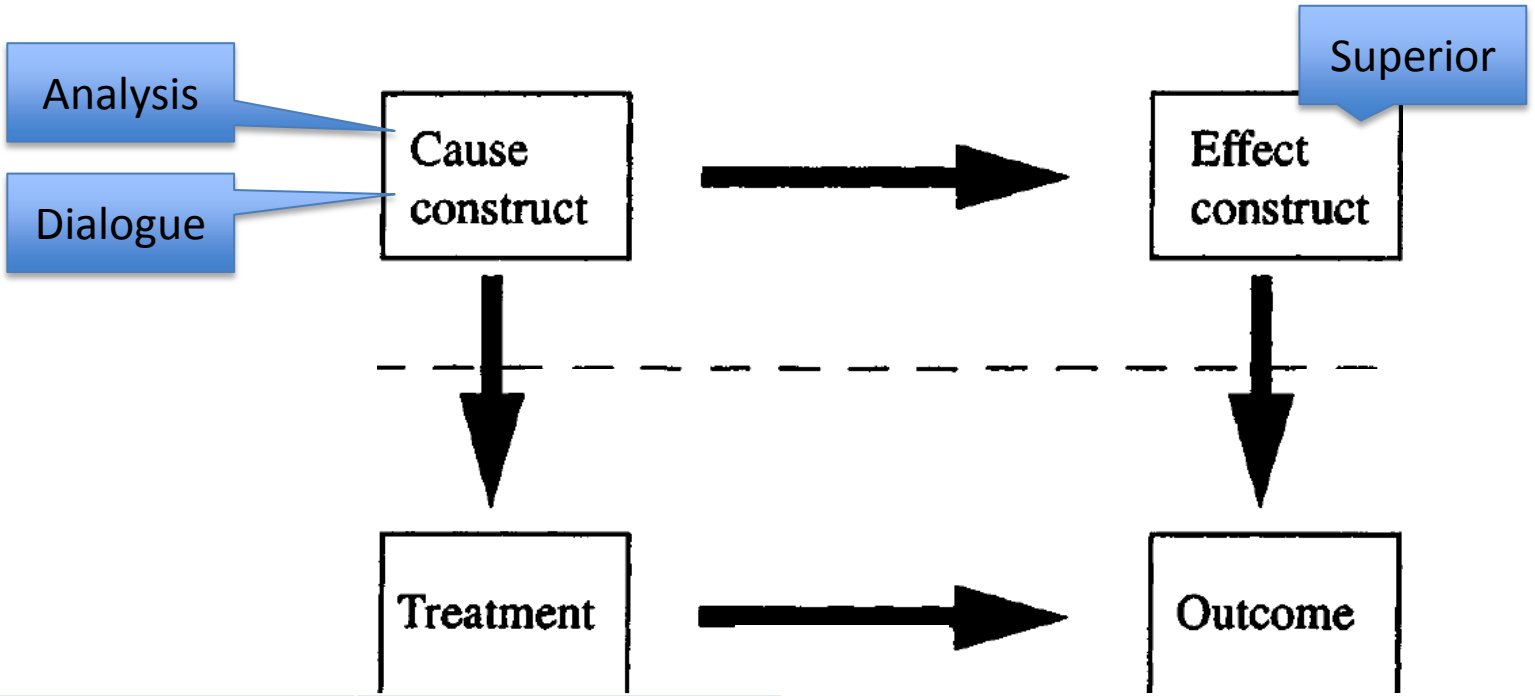
Jeffries et al (2001): Extreme Programming Installed. Addison-Wesley.

Example: Requirements Prioritization Study

- Research Questions:
  Which prioritization technique is superior?
  - Pair-wise comparisons (PWC)
  - Planning game (PG)
- Hypotheses
  - $H_0$1: Time to conclude prioritization is equal
  - $H_0$2: The ease of use of the techniques is equal
  - $H_0$3: The accuracy of the prioritization results is equal
  - H_A1: Time to conclude prioritization is NOT equal
  - H_A1: The ease of use of the techniques is NOT equal
  - H_A1: The accuracy of the prioritization results is NOT equal

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Example: Requirements Prioritization Study



8 Requirements	16 Requirements
PWC, then PG	PWC, then PG
PG, then PWC	PG, then PWC

Which of the two requirements is more valuable to you?

Alarm <<<< <<< << < = > >> >>> >>>> Timer

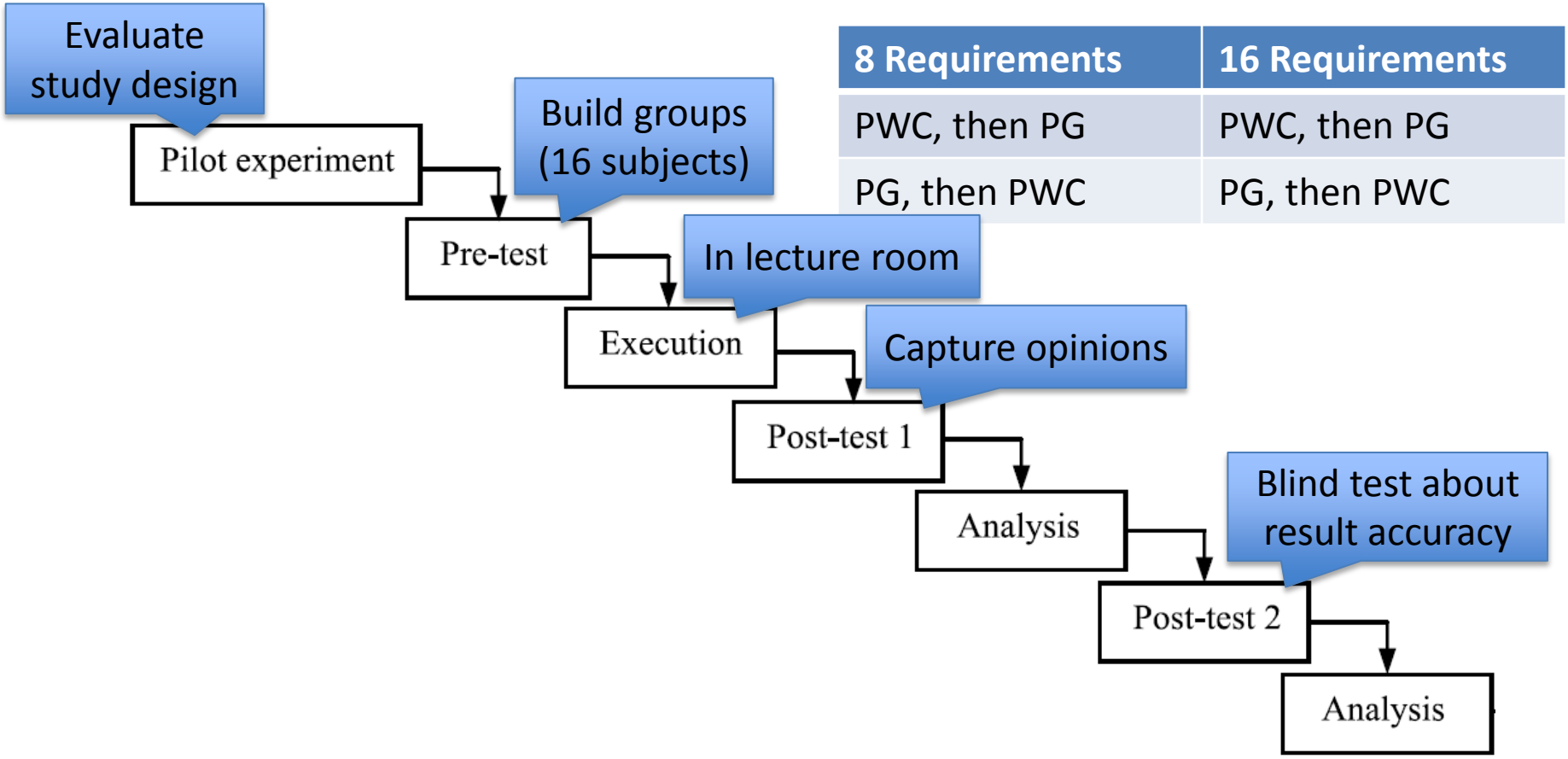
WAP <<<< <<< << < = > >> >>> >>>> SMS

…

Which result set is more accurate?

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Conducted Activities

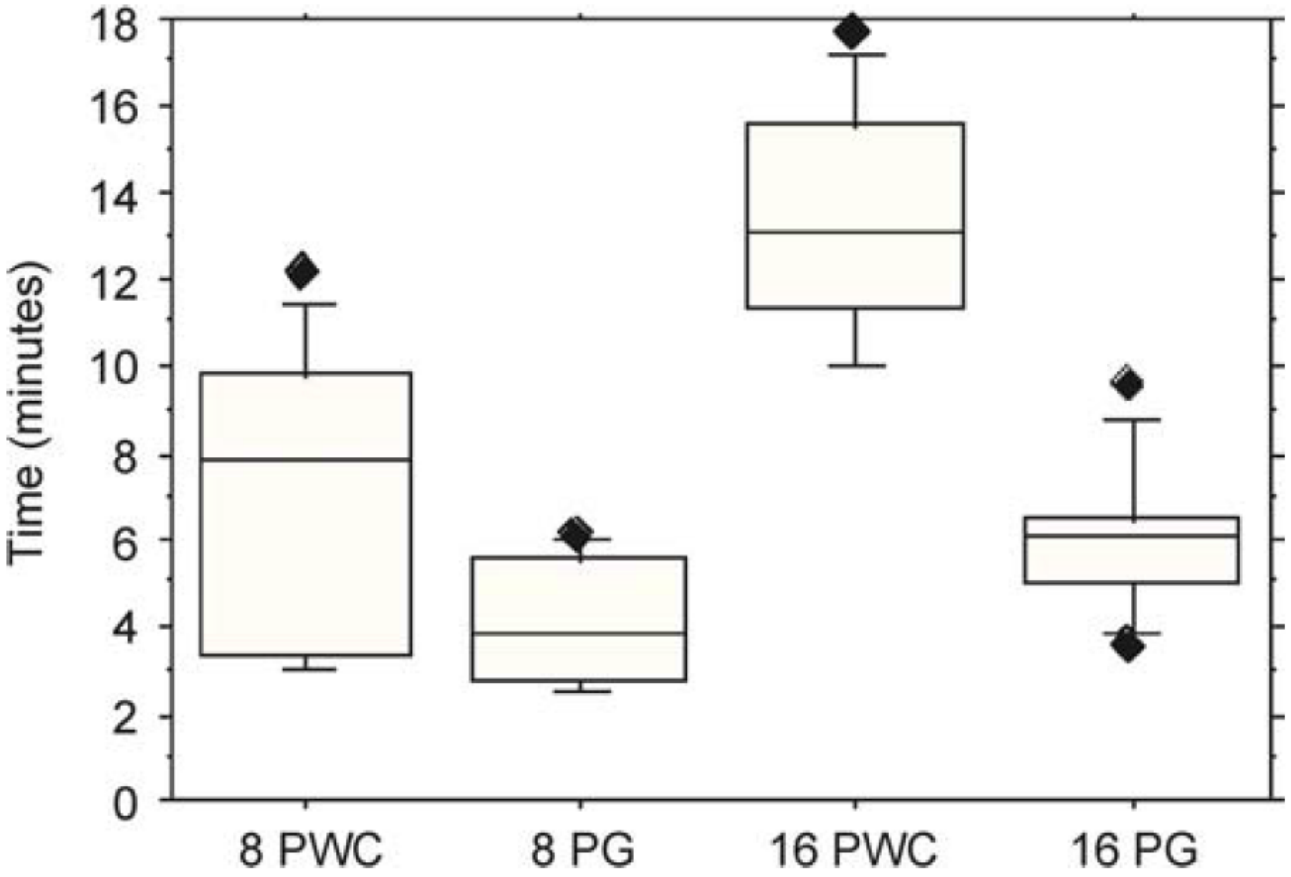


8 Requirements	16 Requirements
PWC, then PG	PWC, then PG
PG, then PWC	PG, then PWC

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Example: Requirements Prioritization Study

- Results: time



Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Example: Requirements Prioritization Study

- Results: accuracy

Nbr of requirements	Criteria	Favour PG	Equal	Favour PWC
8	Value	6	2	0
	Price	1	3	4
16	Value	4	1	3
	Price	4	2	2
Total		15	8	9
Total %		**47%**	**25%**	**28%**

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

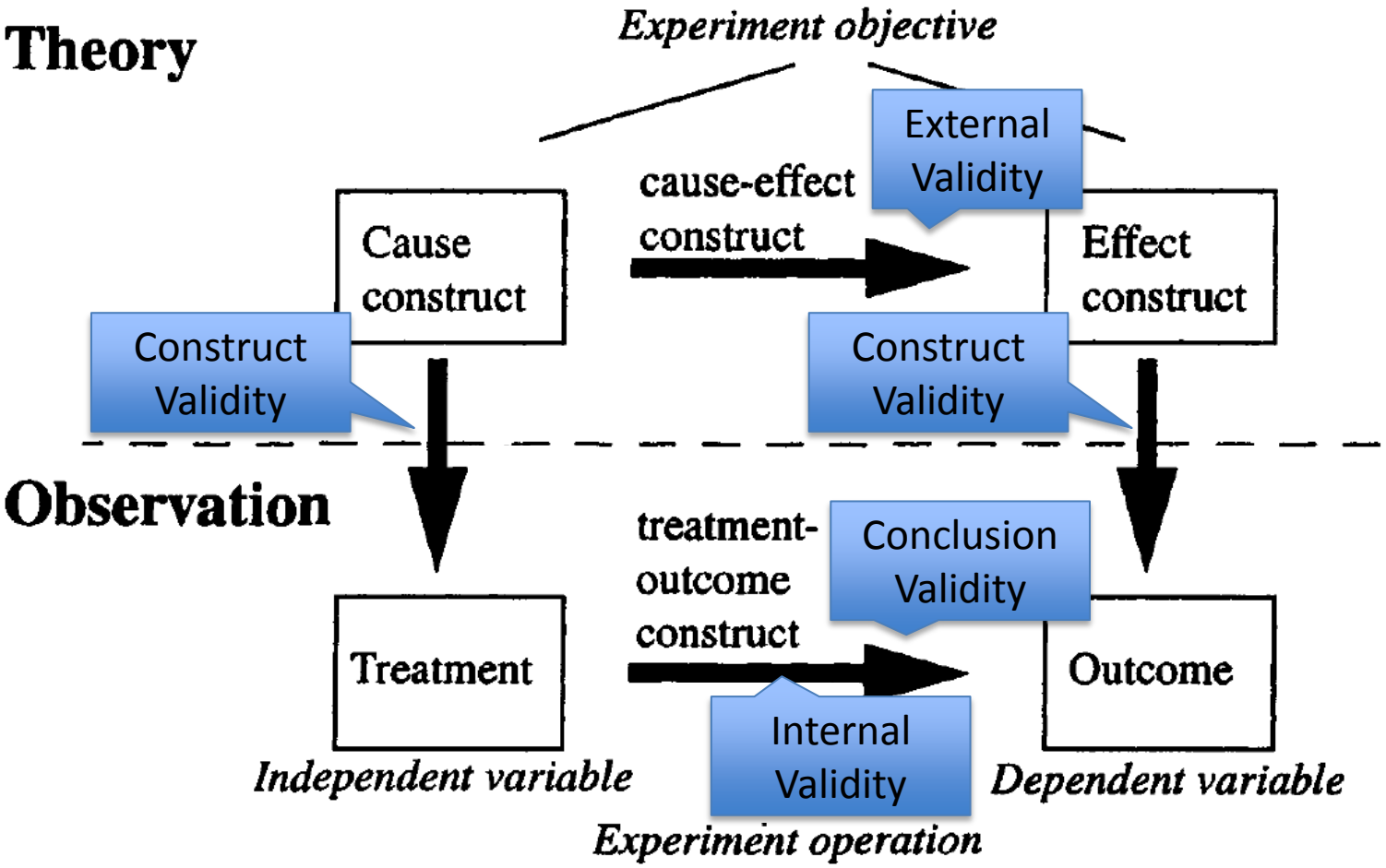
Example: Requirements Prioritization Study

- Results: ease of use

Nbr of requirements	PG much easier	Easier	Equally easy	Easier	PWC much easier
8	4	3	1	0	0
16	4	1	2	1	0
Total	8	4	3	1	0
Total %	**50%**	**25%**	**19%**	**6%**	**0%**

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

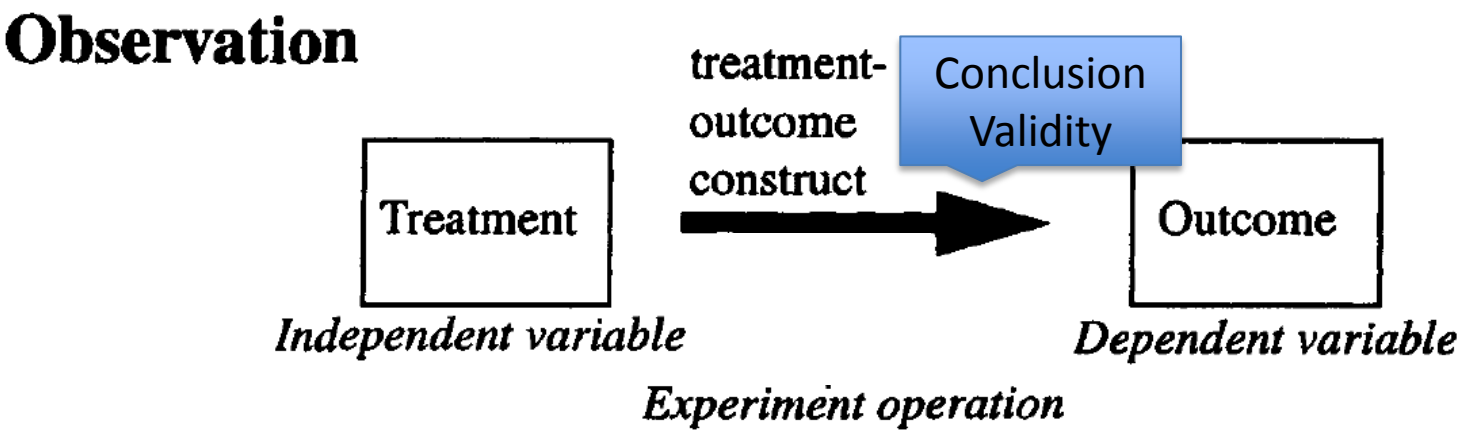
Validity Threat Categories



Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Conclusion Validity

- Are we observing a relationship?
  - Is the relationship between treatment and outcome statistically significant?



Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Conclusion Validity Threats

Example: Requirements Prioritization Study

- Mitigated
  - Robust statistical techniques are used: non-parametric tests because data was not normally distributed
  - High measurement realiability: wording and work support were piloted and adjusted
- Remaining
  - Statistical power is low: only 16 subjects used
  - Measurement objectivity: ease of use and accuracy are subjective

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

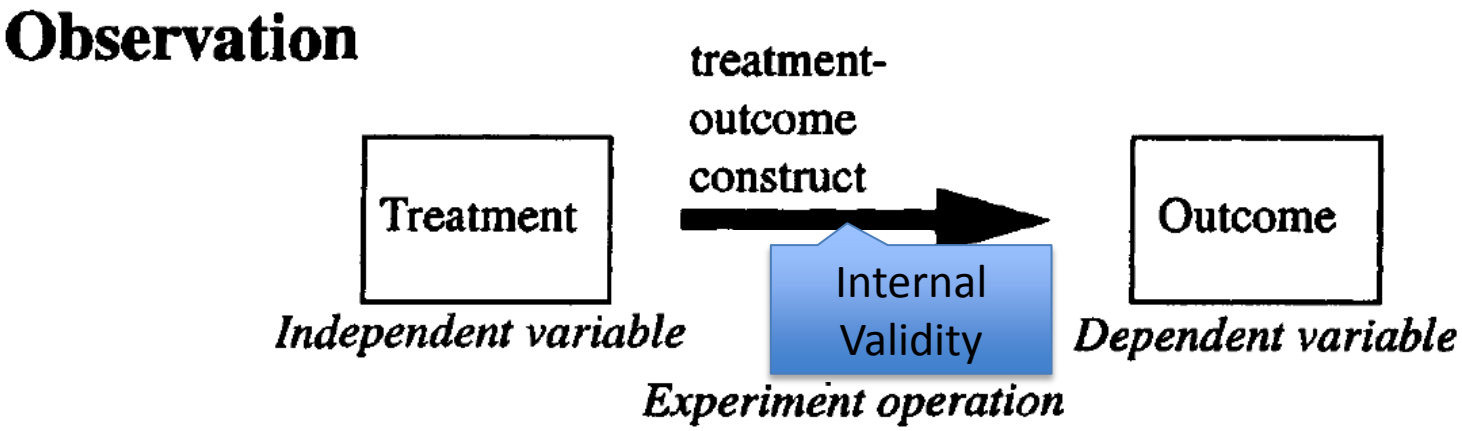
Typical Conclusion Validity Threats

- Low statistical power
- Violated assumption of statistical tests
- Fishing and the error rate
  - The researchers may influence the result by looking for a specific outcome
- Reliability of measures
- Reliability of treatment implementation
- Random irrelevancies in experimental setting
  - Elements outside the experimental setting may disturb the results
- Random heterogeneity of subjects
  - Variation due to individual differences is larger than due to the treatment.

Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Internal Validity

- Does the treatment affect the outcome?
  - Is the relationship between treatment and outcome causal?



Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Internal Validity Threats

Example: Requirements Prioritization Study

- Mitigated
  - Consistency of the results does not differ: learning effects did not affect the experiment
  - Consistency index for PWC did not change: subjects did not get tired during the experiment (concentration not affected)
  - Time used and consistency of the results do not correlate: there was no group pressure to rush
- Remaining
  - (none reported)

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Typical Internal Validity Threats

- History
- Maturation
- Testing
  - Subjects learn and improve when they know the results
- Instrumentation
  - Documents and forms can hint at desired results
- Statistical regression
  - Subjects are grouped according to their pre-experiment performance
- Selection
  - Subjects are grouped according to their natural performance
- Mortality
  - Subject that drop out may have particular characteristics
- Ambiguity about direction of causal influence
  - (A correlates with B) is not the same as (A implies B)
- Diffusion or imitation of treatments
  - One group may imitate the other group
- Compensatory equalization of treatments
  - Compensation of the non-treatment affects the behavior of the control group
- Compensatory rivalry
  - Subjects receiving less desirable treatments may reduce the expected outcome
- Resentful demoralization
  - Subjects receiving less desirable treatments may give up

Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

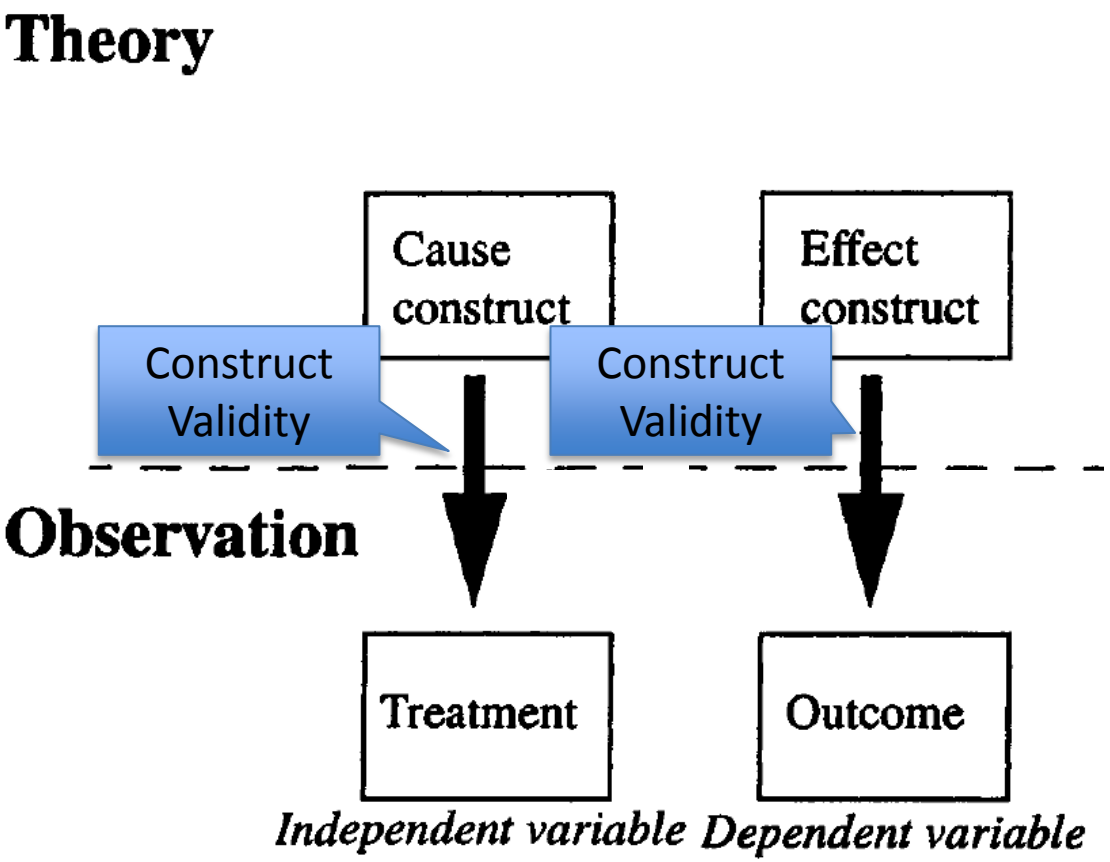
Correlation and Causation

- Examples of wrong causation
  - The more firemen fighting a fire, the bigger the fire is observed to be. Therefore firemen cause fire.
  - Increased pressure is associated with increased temperature. Therefore pressure causes temperature.
  - As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream causes drowning.
  - With a decrease in the number of pirates, there has been an increase in global warming over the same period. Therefore, global warming is caused by a lack of pirates.

- Determining causation:
  - Equality of groups
  - Comparison of treatment vs. no treatment
  - Confirming absence of internal threats
  - Statistical testing of likeliness of the treatment having a causal effect on the disease (e.g. P-value).

Construct Validity

- Do we observe the right things?
  - Does the treatment reflect the cause construct?
  - Does the outcome reflect the effect construct?



Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Construct Validity Threats

Example: Requirements Prioritization Study

- Remaining
  - Judgment of accuracy may be flawed: subjects had their own interpretation of the requirements (requirements initially not explained)

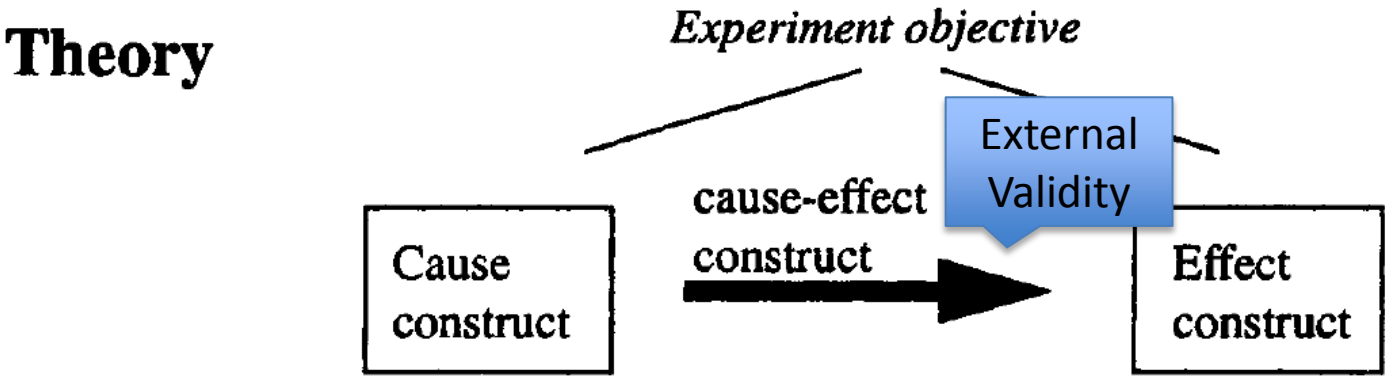
Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Typical Construct Validity Threats

- Inadequate preoperational explication of constructs
- Mono-operation bias
  - A single independent variable, case, subject or treatment may under-represent cause constructs
- Mono-method bias
  - A single measurement variable may under-represent effect constructs
- Confounding constructs and levels of constructs
  - Variables are sometimes not binary, but have more values
- Interaction of different treatments
  - Treatments from different studies may interact
- Interaction of testing and treatment
  - The testing itself may make subjects more sensitive to the treatment
- Restricted generalizability across constructs
  - A construct may be interpreted in inconsistent ways
- Hypothesis guessing
  - Subjects may guess the intended results of the study
- Evaluation apprehension
  - People try to look better when being evaluated (Hawthorne effect)
- Experimenter expectancies
  - Experimenter can bias the study unconsciously

Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

External Validity



- Can the results be generalized outside the scope of the study?

Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

External Validity Threats

Example: Requirements Prioritization Study

- Remaining
  - Subjects are sampled from software engineering PhD students. One study shows, however, that if a student experiment shows that one technique is better than another it is rather unlikely that professionals would come to the opposite conclusion.
  - Number of prioritised requirements is few.
  - The requirements used in this experiment are rather independent.

Karlsson et al (2006): Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Engineering 12:3-33.

Typical External Validity Threats

- Interaction of selection and treatment
  - Having a subject population not representative of the population we want to generalize
- Interaction of setting and treatment
  - Not having the experimental setting or material representative of, for example, industrial practice
- Interaction of history and treatment
  - The experiment is conducted on a special time or day which affects the results

Wohlin et al (2000): Experimentation in Software Engineering – An Introduction. Kluwer.

Contents

Understand

- Threats to validity

Agenda

- Threats to Validity in Experimentation
- Threats to Validity in Other Research Methods

Exercise

- Goal:
  - Understand specific kinds of validity threats
- Research types:
  - SLR
  - Case study
  - Survey
- Tasks
  - Study research report
  - Summarize study design and results
  - Discuss reported threats to validity and their differences with those of an experimental setting

Contents

Understand

- Threats to validity

Agenda

- Threats to Validity in Experimentation
- Threats to Validity in Other Research Methods