Empirical Software Engineering (EMSE) - A Vehicle for Evaluating Products and Processes

Dietmar Winkler

Vienna University of Technology
Institute of Software Technology and Interactive Systems

dietmar.winkler@qse.ifs.tuwien.ac.at
http://qse.ifs.tuwien.ac.at

Motivation

- A major goal in software engineering is the delivery of high-quality software solutions.
- The construction of software products requires professional approaches, e.g., software processes (e.g., Life-Cycle Model, V-Modell XT, Scrum).
- Methods support engineers in constructing and evaluating software products.
  - Constructive approaches, e.g., Model-Driven Development, Test-Driven Development, and Pair Programming to create new software products.
  - Analytical approaches, e.g., inspection and testing to assess product and process quality.
- Increasing product quality (e.g. less defects), project and process performance (faster delivery of products) requires the application of improved methods and tools.

Questions

→ How can we evaluate and assess improved methods and processes?

→ How can we measure process / product attributes in general?

→ How can we conduct an empirical study?

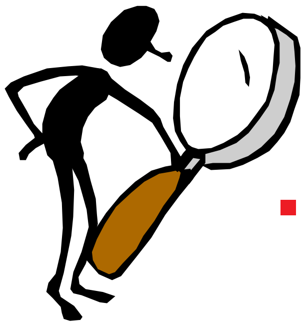
Why Empirical Studies?

- New software development technologies come up frequently, e.g. tools, methods: Why should we invest in those technologies?

- In other disciplines, technology evaluation is a pre-requisite (e.g., medicine), … but not in software engineering…
  Often intuition: “I believe that my method is better than XYZ”?

Examples

- Product evaluation, e.g., prototyping.
- Process evaluation
  - Prototypes are not possible (simulation based on models).
  - A process is just a description until it is used by people.



- Important for research: experimentation is mandatory in other disciplines (e.g., medicine, physics, etc.)

- Experimentation provides a systematic, disciplined, quantifiable and controlled way of evaluating human-based activities.

Goals and Benefits of Empirical Studies

The purpose of a study is

- to explore ...
  - finding out what's happening
  - seeking for new insights
  - asking questions and to find answers

  Measurement: usually qualitative

Determine what you want to learn

- to describe ...
  - portray accurate profile of situations, events, projects, technologies

  Measurement: quantitative/qualitative

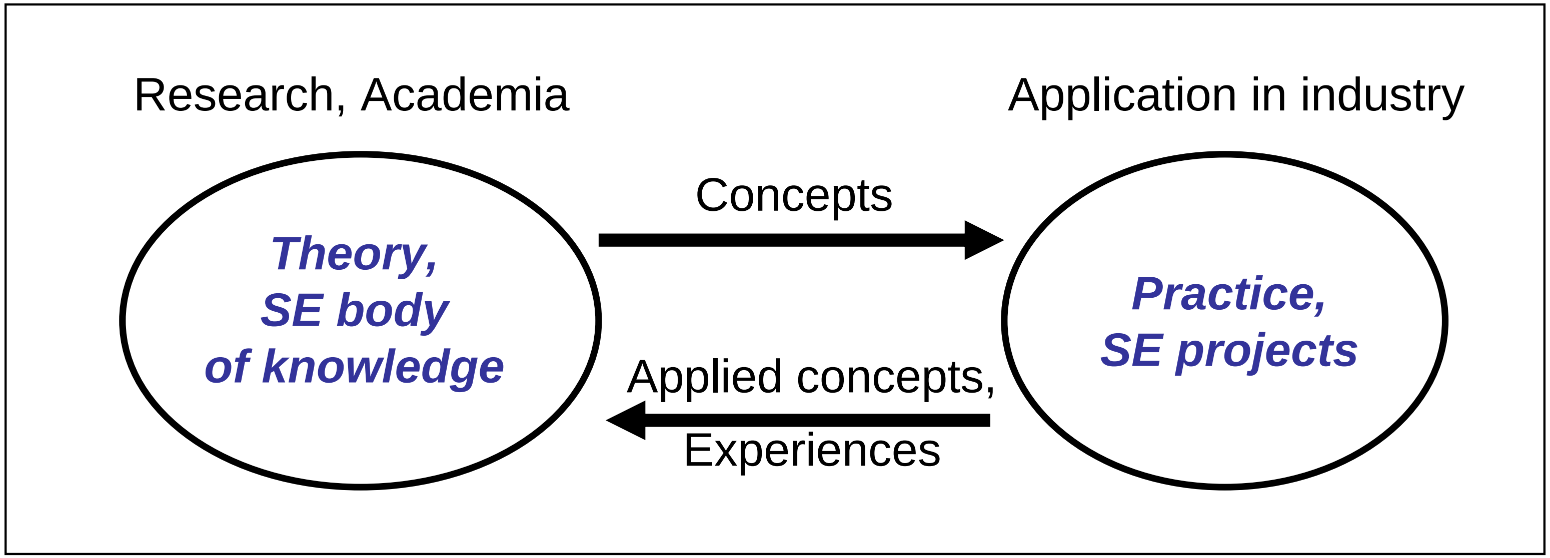
- to explain ...
  - seek explanation of a situation/problem, usual in the form of causal relationships

  Measurement: quantitative/qualitative

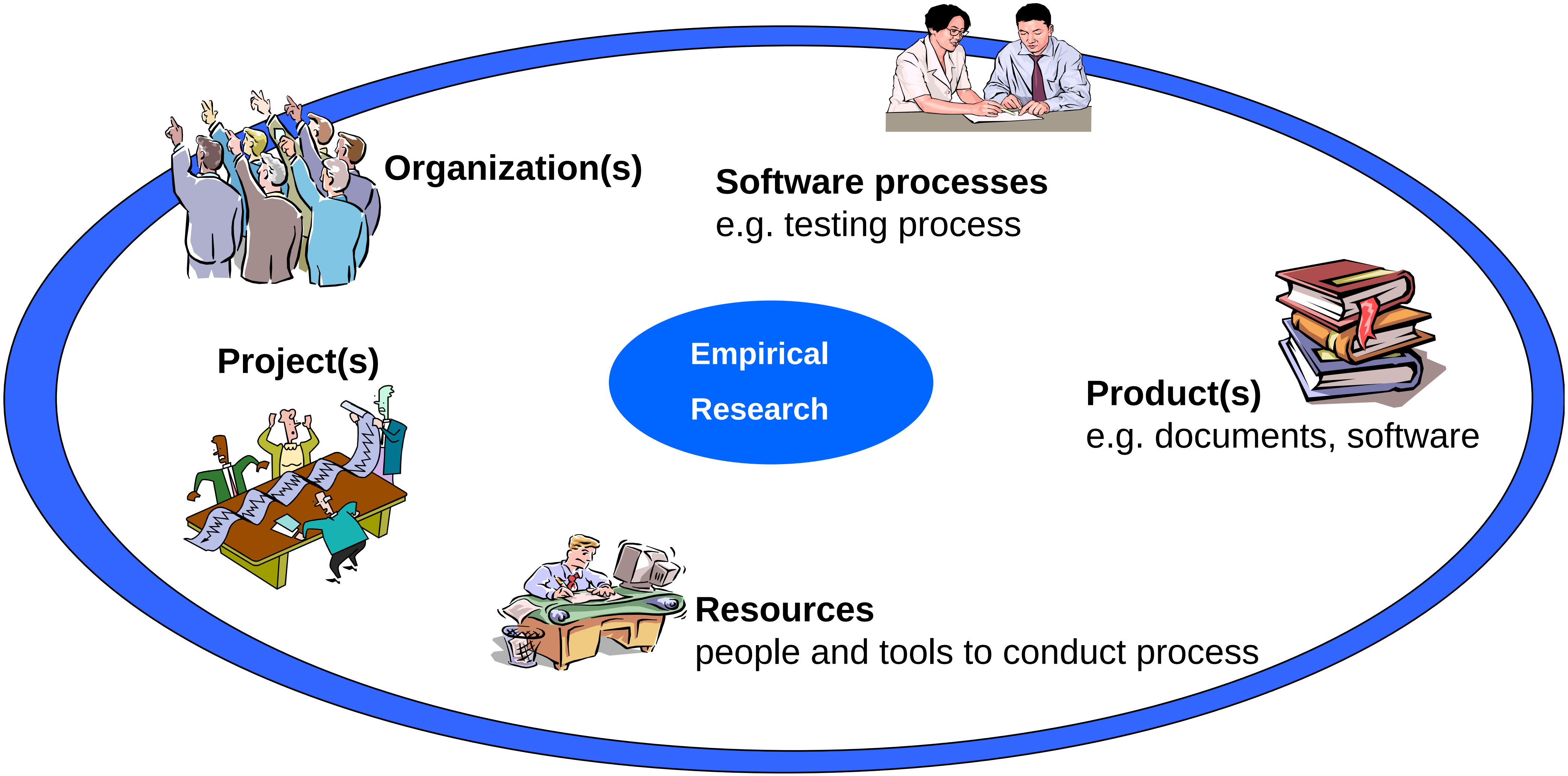
- … relationships, differences, and changes

Empirical Studies in SE

- Conducting empirical studies is research to improve Software Engineering Practice.
  - Apply theoretical concepts in SE practice.
  - Add experiences on the appliance to the SE 'body of knowledge'
  - Improve processes, methods and tools (SPPI approach).
  - Verify theories and models.



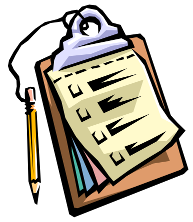
Objects of Empirical Research



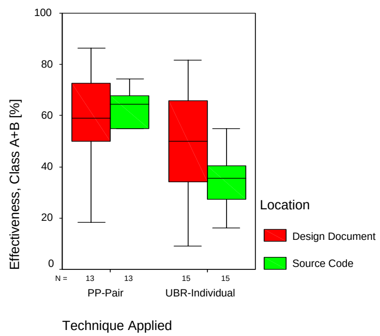
Some Basic Concepts

- Measurement
  - is the process of capturing data which are connected to real-world attributes to describe them.
  - Why is measurement important?

- Data Collection
  - Collection of qualitative / quantitative data according to research questions.
  - How can we collect data?



- Data Analysis
  - Analyzing the results according to the research questions
  - Statistical tests to report significant results.
  - Which information can we derive from collected data?



Measurement

- **Quotes**:
  - "You can't manage what you can't measure", Tom DeMarco
  - "What is not measurable make measurable", Galileo Galilei

- **Objectives**:
  - One objective of science is to find ways to measure attributes of entities we are interested in.
  - Measurement makes concepts more visible and thus more understandable and controllable.

- **Definition**
  - Measurement is the process by which numbers or symbols are mapped to attributes of entities in the real world in such a way as to describe them according to clearly defined rules.

Measurement (Examples)

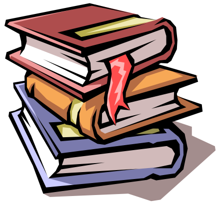
Process

effort → 10 weeks

person days spent from start to end

- Examples: Development process (V-Modell XT), Testing Process, Inspection, ..

Product



size → 700 LOCs

No. Lines of Code

- Examples: Design Specification (No of pages), Test Suite (number of test cases), Module (LoCs)

Resources

experience → high

>10 projects

- Examples: Project management experience, Testing experience, Design / Architecture experience.

Selected Types of Measures

- Direct vs. Indirect Measures:
  - Direct: obtaining values direct from the study object (e.g., duration, effort)
  - Indirect: calculated values based on various attributes (e.g., efficiency of defect detection = number of defects per time interval)

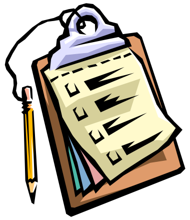
- Objective vs. Subjective Measures:
  - Objective: no judgment of the measurement value (e.g., LoC, delivery date)
  - Subjective: reflect judgment of the measurer, depending on the viewpoint (e.g., subject defect estimation, questionnaires)

- Quantitative vs. Qualitative data:
  - Quantitative: data expressed as numbers (e.g., data obtained through measurement, statistics)
  - Qualitative: data expressed as word and pictures (e.g., interviews, interpretation)

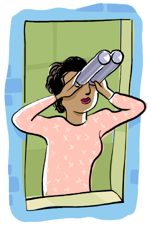
Data Collection

- Measurement
  focuses on products, processes (typically quantitative data collection)

- Interviews
  based on information obtained from individuals persons or groups (typically qualitative data)



- Questionnaires
  set of questions to obtain information from individuals, e.g., experience, feedback; (typically used in surveys)



- Observation
  selection, recording, and encoding of a set of natural behaviours or other naturally occurring phenomena (typically used in case studies)

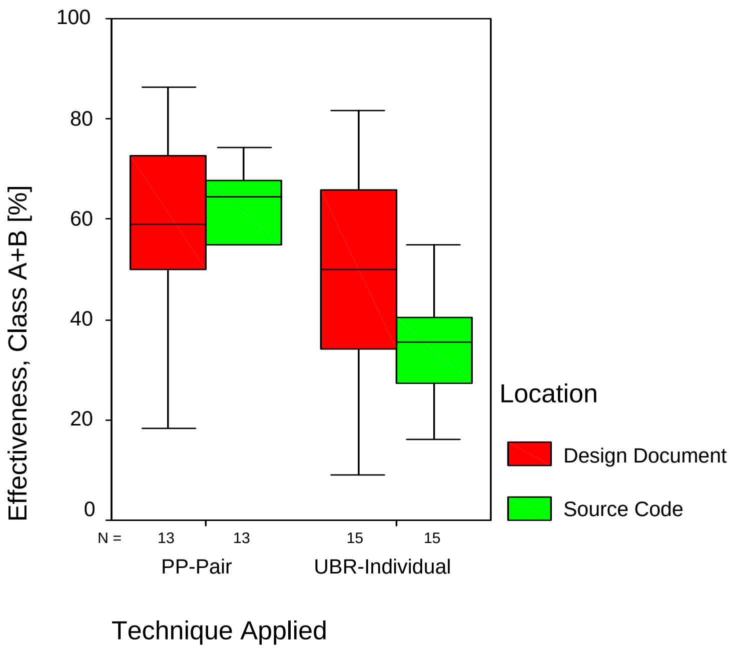
Data Analysis

Purposes of quantitative data analysis

- Describing a population (descriptive statistics)
- Exploring differences between groups (Hypothesis Testing)

Examples:

- Minimum, Mean, Maximum, Standard Deviation.
- Visualization, Statistical Tests to test Hypothesis.



Statistical Tests

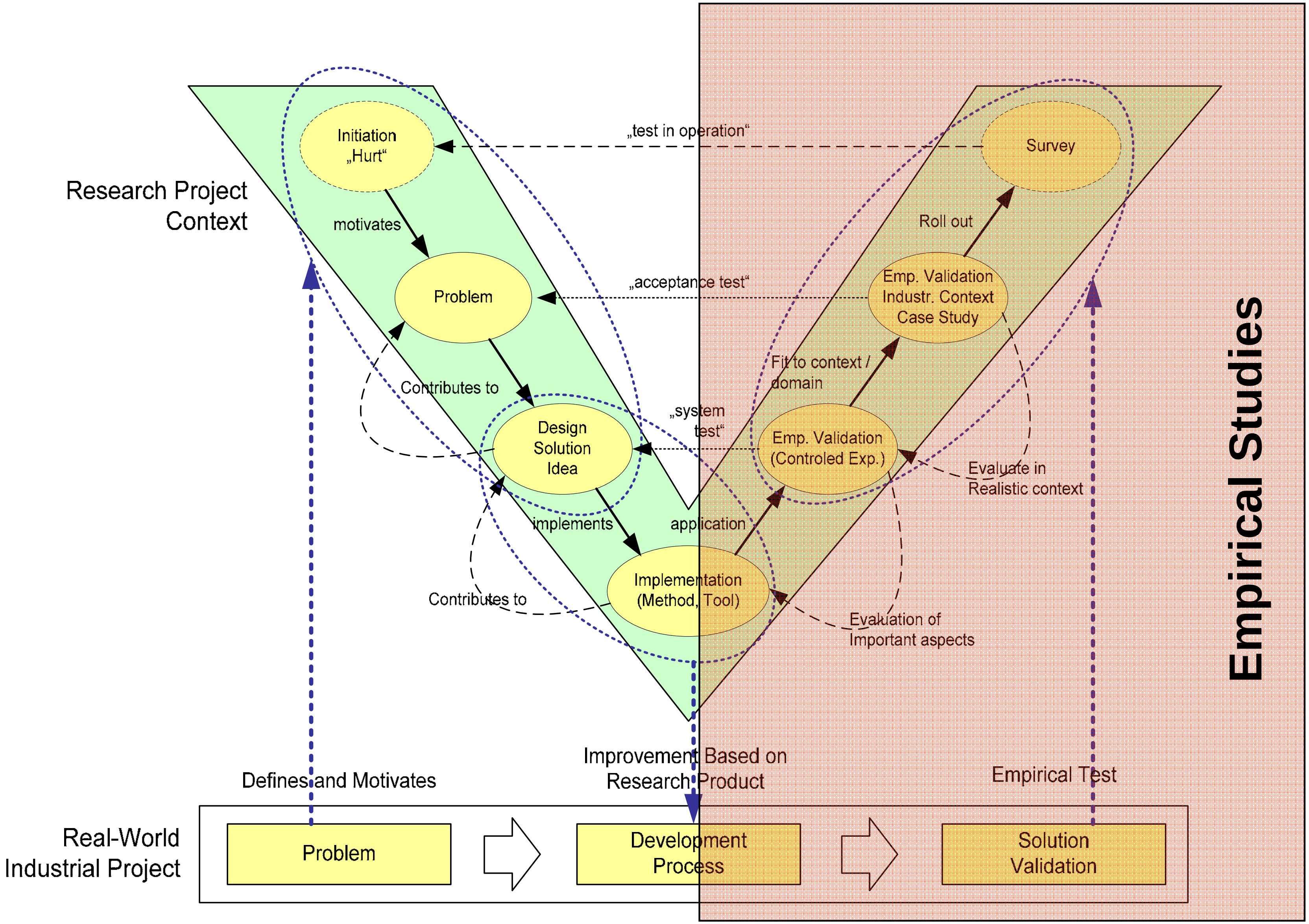
	Location	PP-Pair	UBR-Individuals	P-value
Mean	DD+SC	56.3	40.3	0.013 (S)
	DD	56.3	47.3	0.212 (-)
	SC	56.3	35.3	0.004 (S)
Std.Dev	DD+SC	20.6	13.6	-
	DD	26.7	20.6	-
	SC	17.9	11.4	-

Classification of Empirical Studies

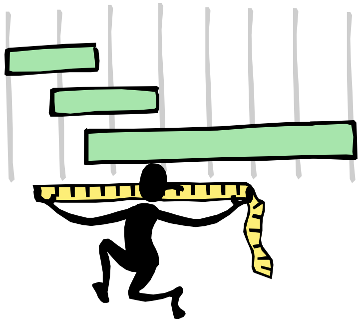
- Empirical studies provide a systematic, disciplined, quantifiable and controlled way of evaluating human-based activities.
- Empirical studies are important for scientific work to generate knowledge of products, processes and resources (“V-model” of empirical research).
- Empirical methods are important techniques for software quality improvement.
- Different study strategies aim at focusing on individual steps of product / process progress (e.g., laboratory evaluation and simulation, organization case studies, cross-company surveys etc.)

- Different Empirical Strategies:
  - Surveys
  - Case Studies
  - Controlled Experiments

Empirical Studies in the Context of the V-Model



Different Empirical Strategies



Controlled Experiments

- Measuring the effects of one or more variable(s) on other variable(s).
- Detailed investigation in controlled conditions (relevant variables can be manipulated directly, precisely and systematically).

Case Studies

- Development of detailed, intensive knowledge about a single case or of a small number of related cases.
- Detailed investigation in typical conditions.

Surveys

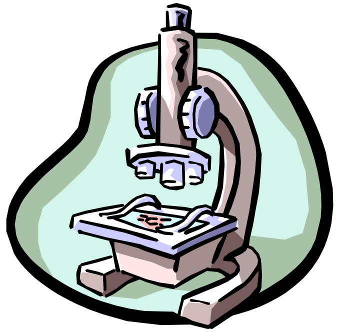
- Collection of information in standardized form within groups of people or projects.
- Usually performed retrospectively.
- The use of a technique/tool has already taken place; relationships and outcomes should be documented.

Empirical Strategies

- Controlled Experiment:
  - laboratory environment.
  - an operation is carried out under controlled conditions.
  - manipulate one or more variables and keep all other variables at fixed levels.
- Case Study:
  - monitoring projects or activities.
  - data collection for a specific purpose.
  - observational study.
- Survey:
  - investigation performed in retrospect.
  - interviews and questionnaires.

Strategy	**Quantitative** (data expressed as numbers)	**Qualitative** (data expressed as words or pictures)	**Study Effort** (always depends on context and research topic)
Experiment	X		(very) high
Case Study	X	X	Medium
Survey	X	X	Low/Medium

Controlled Experiment: Fact Sheet



Purpose:

- Detailed investigation in controlled conditions (relevant variables can be manipulated directly, precisely and systematically)

When select an experiment?

- When appropriate: control on who is using which technology, when, where and under which conditions.
- Level of control: high
- Data collection: process and product measurement, questionnaires
- Data analysis: statistics, comparison of groups, etc.

- Pro's: help establishing causal relationships, confirm theories.
- Con's: representative experiment setting?
  Challenging to plan in a real-world environment.
  Application in industrial context requires compromises.

See example later in this lecture

Case Study: Fact Sheet

Purpose:

- Development of detailed, intensive knowledge about a single case or of a small number of related cases.
- Detailed investigation in typical conditions.

When select a Case Study?

- When appropriate: change (new technology) within a development process, we want to assess a change in a typical situation. Project monitoring.
- Level of control: medium
- Data collection: product and process measurement, questionnaires, interviews.
- Data analysis: compare case study results to a baseline (sister project, company baseline).

- Pro's: applicable to real world projects, help answering why and how questions, provide qualitative insights.
- Con's: difficult to implement a case study design, analysis of results is subjective

Survey: Fact Sheet

Purpose:

- A retrospective study of a situation to try to document relationships outcomes.

When select a Survey?

- When appropriate: for early exploratory analysis.
  Technology change implemented across a large number of projects, description of results, influence factors.
- Level of control: low
- Data collection: questionnaires, interviews
- Data analysis: comparing different populations among respondents, association and trend analysis, consistency of scores.

- Pro's: generalization of results is usually easier (than case study), applicable in practice.
- Con's: little control of variables, questionnaire design is difficult (validity, reliability), execution is often time consuming (interviews).

Typical Survey Type

State-of-the-art Surveys

- Ask people on state-of-the-practice / best practices.
  - Inside an organization: people, departments, business units.
  - Over organizations: people with a specific function (e.g. QA, engineer), people in specific departments.
  - Example: Which software processes do you use in your organization?

Literature Surveys

- Analyze existing literature (papers, books, notes) to determine the state-of-the-art, best practices on a topic.
- Common practice for research!

Trend Surveys

- Evaluate demand of particular products or services and predict their future.
  - Conducted by institutes like Ovum, Gartner & IDC.
  - Also by asking people in organization.

Selecting an Empirical Strategy

How to select the appropriate strategy for a study:

Purpose of study

- Exploratory, descriptive or confirmatory.
- Questions concerning what, how, how many, where, for whom.

Degree of control

- Possibility to ‘arrange’ the real world.
- Required versus possible degree of control.

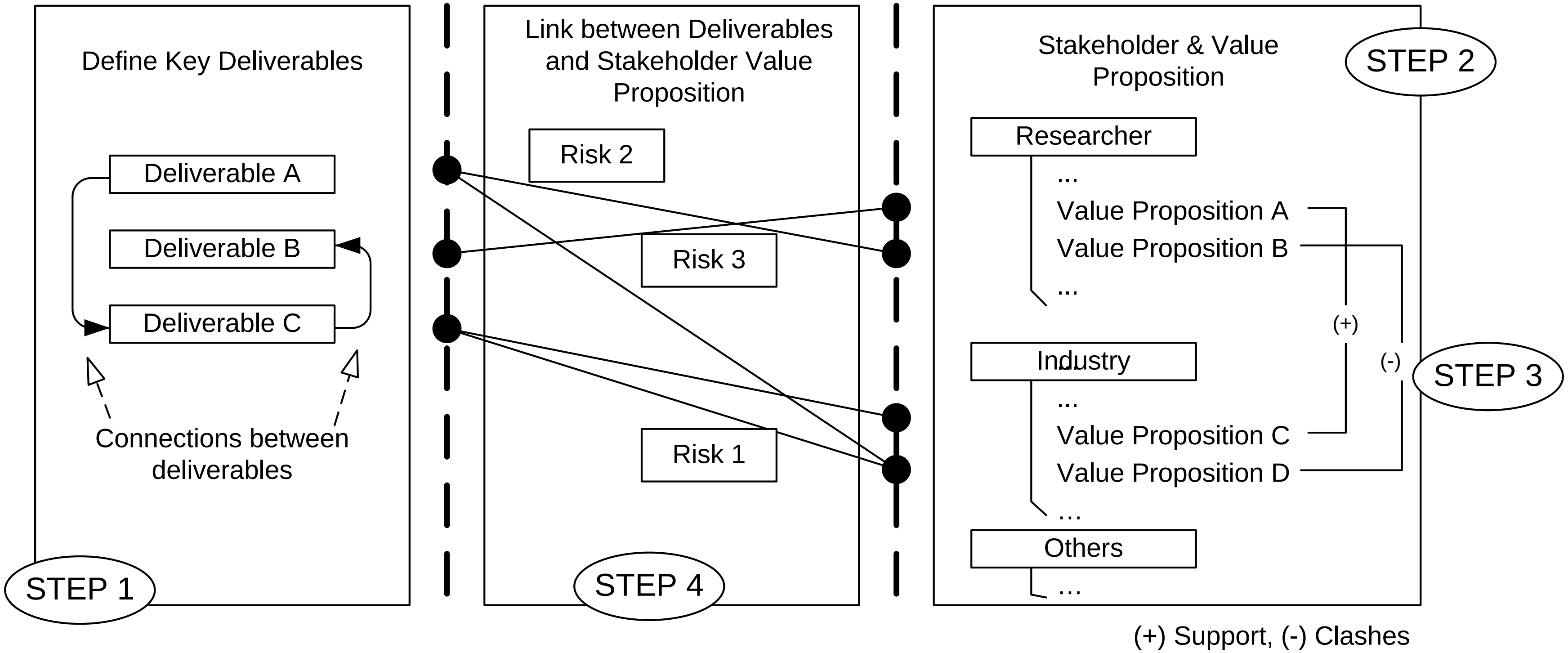
Cost / Effort

- The relative costs for doing a study;
  e.g. costs for doing experiments are considered as being high.

Risk

- Probability that study might fail and its consequence.

Value Based Empirical Research Plan Evaluation



- **Step 1**: Characterization of key deliverables and dependencies between them.
- **Step 2**: Elicitation of key stakeholders (industry and academia) and their value propositions.
- **Step 3**: Identification of significant support (+) and conflicts (-) between stakeholder value propositions.
- **Step 4**: Linking of deliverables to stakeholder propositions; risk analysis: e.g., unaddressed stakeholder win conditions, necessary additional activities.

Biffl et al, 2007

Summary

- Experimentation provides a systematic, disciplined, quantifiable and controlled way of evaluating human-based activities.
- The purpose of a study is to explore, to describe, and to explain relationships, differences, changes of products, processes, and resources.
- Measurement provides quantitative and qualitative data of the study object in an objective and/or subjective way. Measures can be collected directly (e.g., effort and defects) or indirectly (e.g., number of defects per hour = efficiency).
- Data collection approaches are basic elements of empirical studies (e.g. measurement, interviews, questionnaires, observation).
- Data analysis describes data of the study, relationships between different entities, etc. Statistical tests are used to falsify hypothesis.
- Main study strategies are controlled experiments, case studies, and surveys.
- VBER provides a systematic planning of empirical studies including key deliverables of the study, involved stakeholders and their value proposition and the link between deliverables and stakeholder value.

Empirical Software Engineering Processes
An Example

Dietmar Winkler

Vienna University of Technology
Institute of Software Technology and Interactive Systems

dietmar.winkler@qse.ifs.tuwien.ac.at
http://qse.ifs.tuwien.ac.at

Motivation

- Experimentation provides a systematic, disciplined, quantifiable and controlled way of evaluating human-based activities.
- The selection of the study strategy depends on the
  - purpose of the study (exploratory, descriptive or confirmatory),
  - the degree of control (high, medium, low),
  - cost/effort for study preparation, execution and analysis,
  - and possible risks.
- Different Study Strategies: Controlled Experiments, Case Studies, Surveys.
- To handle complex study processes, researchers have to follow a pre-defined sequence of steps (study process)

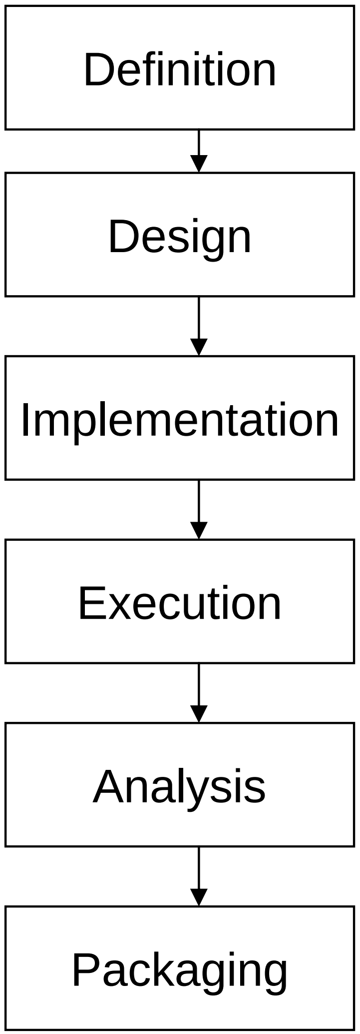
Questions

→ What major steps must be considered in conducting an empirical study?

→ What are the major issues to control.

Controlled Experiment – Basic Process

An overview on the high level process



- Definition: Determine study goal(s) and research hypothesis(es). Select type of empirical study to be employed.
- Design: Operationalize study goal(s) and hypotheses. Make study plan: what needs to be done by whom and when.
- Implementation: Prepare material required to conduct the study.
- Execution: Run study according to plan and collect required data.
- Analysis: Analyze collected data to answer operationalized study goal and hypotheses
- Packaging: Report your study so that external parties are able to understand results and the context of the study.

Research Proposal: Content

1. **Introduction and motivation**
   - why is the research relevant.
   - description of issues or points.
2. **Relevant prior work**
   - what is the work based on.
   - what are the other relevant research results.
   - what is the "research gap" that this research contributes to.
   - it is sufficient to refer to main relevant work.
3. **Research Objectives, questions and hypotheses**
   - explicit articulation of the research objectives (higher level goals for the research)
   - explicit definition of the research hypotheses and questions (more specific statement)
4. **Empirical study design and arrangements**
   - overall design of the study.
   - description of study arrangements.
   - description data collection procedures and protocols.
5. **Definition of metrics**
   - definition of metrics used in the study, include a list and definition of most important metrics.
6. **Data analysis methods**
   - description of the methods and techniques used in data analysis.
7. **Validity threats and control**
   - description of potential threats and how they will be mitigated
   - how generalizeable the results are?

Example: Idea & Background

Basic Idea:

- Improving product development applying agile development practices.

Background:

- Pair Programming (PP)
  - is a flexible and constructive approach for software development in short iterations.
  - supports tight customer interaction and frequent requirements changes.
  - focuses on software construction performed by 2 persons sharing a common working environment.
- Analytical Quality Assurance (QA) Activities, e.g., software inspections, testing
  - are sometimes considered as add-on activity in software development (even if time is very short).
  - supports systematic defect detection and product improvement.
- The idea is to bundle the benefits of pair programming and software inspection to improve software products!

Note: this presentation contains a subset of the overall experiment setting

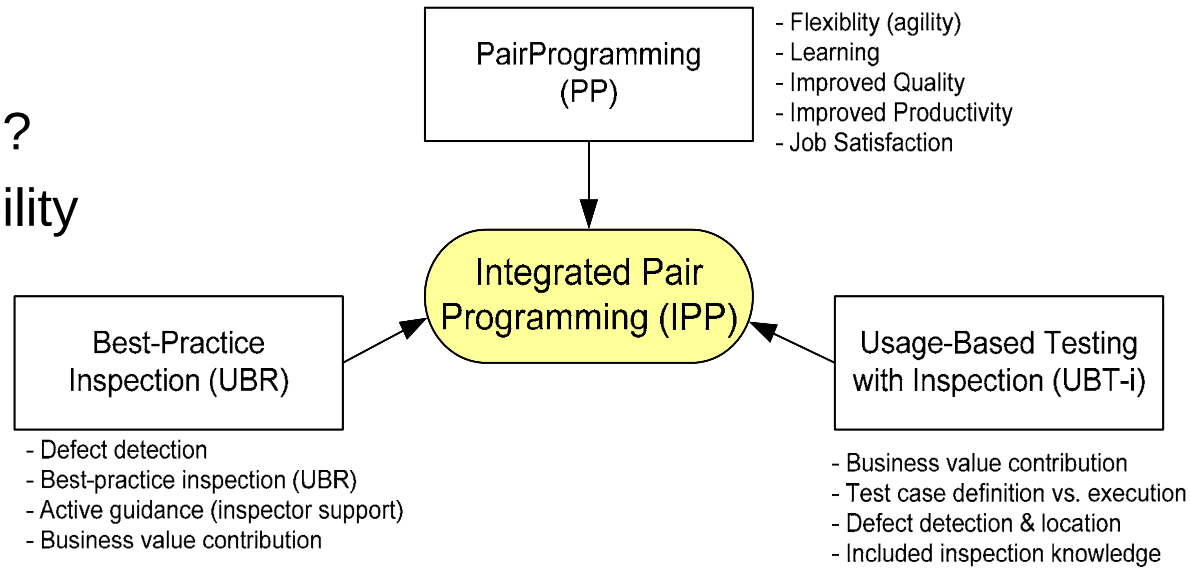
Example: Benefits of the Approach

- In traditional pair programming the observer performs some quality assurance activities, e.g., implicit continuous reviews.
- This implicit quality assurance is not well defined, not traceable and not repeatable.

- Thus, traditional pair programming is not suitable for environments that need well-defined, traceable and repeatable quality assurance (e.g., security-related application domains).

Main Questions:

- How to integrate QA in PP?
- How can we show traceability and repeatability?
- Effects of QA on defect detection?



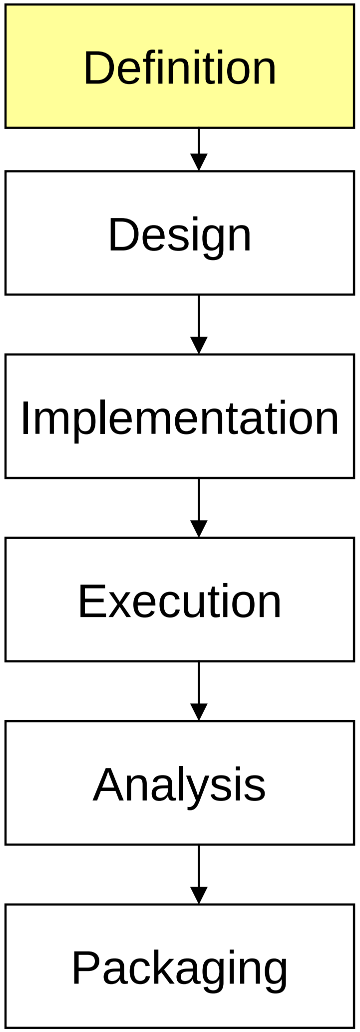
Example:
Idea to conduct an Empirical Study …

- Topic: Integration of Analytical Quality Assurance Methods into Agile Software Construction Practice → "An Integrated Pair Programming Approach" (IPP)

- Type of Study: Controlled Experiment
  - When appropriate: control on who is using which technology, when, where and under which conditions.
  - Level of control: high
  - Data collection: process and product measurement, questionnaires
  - Data analysis: statistics, comparison of groups, etc.

- Research proposal available at: http://www.sbl.tkk.fi/idoese/

Experiment Process: Definition

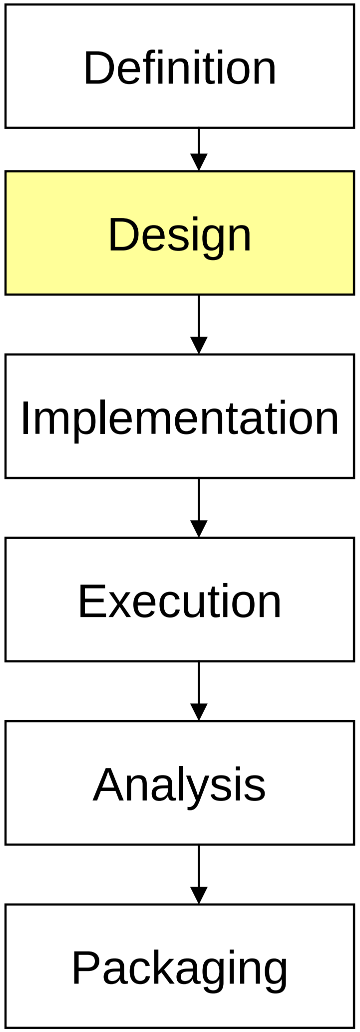


- Determine study goal(s) and research hypothesis(es). Select type of empirical study to be conducted.
- Define Research Objectives:
  - explicit articulation of the research objectives (higher level goals for the research)
  - Example: the new model will increase software product quality.
- Define Hypotheses:
  - explicit definition of the research hypotheses and questions (more specific)
  - Example: Method 1 performs better than method 2, because …

Example: Definition

- Research Objectives
  - Improve product quality bundling constructive (PP) and analytical (inspection) SE & QA approaches.
  - Establish explicit (systematic, traceable and repeatable) QA in agile construction practice (IPP).
- Study Goal:
  - Investigation of Defect Detection Capability of new and traditional approaches.
- Hypothesis:
  - H1.1 Efficiency (IPP) > Efficiency (Inspection)
    Expectation: Bundling benefits of PP and Inspection will increase defect detection efficiency (defect detection over time) significantly in contrast to software inspection.
  - H1.2 Efficiency (IPP) > Efficiency (Inspection) in source code documents
    Expectation: IPP uses a compiler, involvement of “two brains” → IPP will perform better than paper-based solo-inspection.

Experiment Process: Design



- Operationalize study goal(s) and hypothesis(es).
  Make study plan: what needs to be done by whom and when.
- Determine what needs to be observed / measured;
  quantitative and qualitative data.
- Maximize validity of results;
  identify what effects might influence my findings.
- Maximize reliability of the study (to enable replication)
  → documentation of procedures, context, measurements.

Example: Design (1)

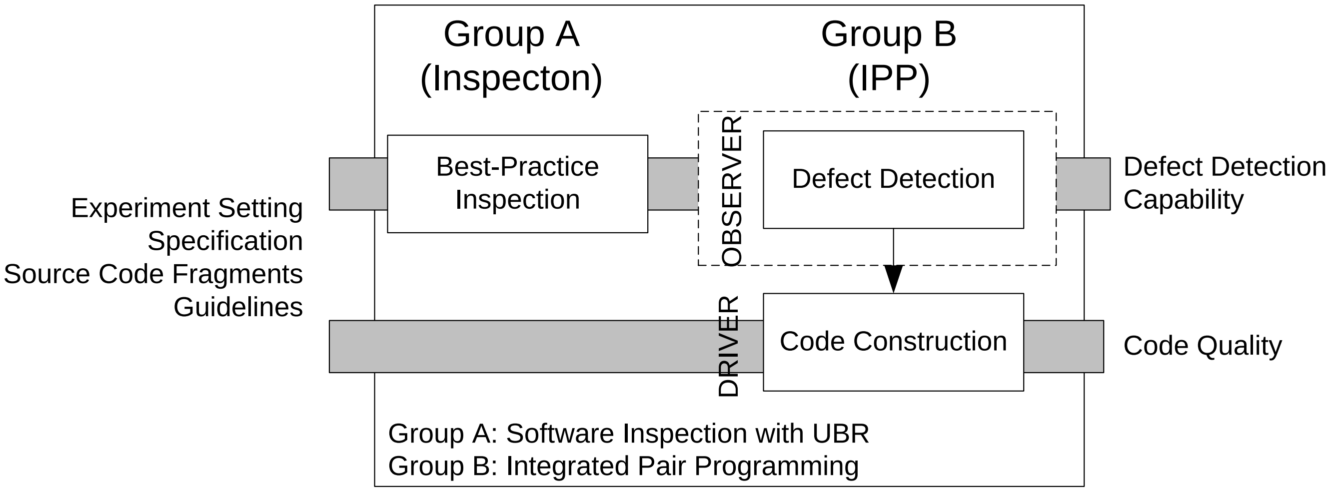
- 5 Basic Steps (Execution Phase of the Study):
  (a) Participant selection, (b) experience collection (questionnaires)
  (c) experiment preparation for participants, (d) study execution in two sessions including feedback questionnaires after every session, and
  (e) data submission.

- Study Material:
  - Scope of the system: Maintenance / evolution process for a commercial application.
  - Application: Taxi-Management system (Dispatcher, Driver) including two system parts (= 2 sessions of the study); well-known application area.
  - Objects: Textual requirements, Prioritized Use Cases, Source Code fragments (partially implemented), Guidelines, Questionnaires.
  - Expert Seeded Defects: 60 defect spread over different document locations (different defect severity classes and types).

- More than 100 overall participants (subjects) in different groups.
  Registration of prior knowledge using questionnaires and other sources.

Example: Design (2)

- Investigation and Comparison of Defect Detection Capability (Effectiveness, Efficiency).
- Direct Measurement:
  - Number of seeded defects.
  - Number of found / matched defects.
  - Defect detection duration (time).
- Indirect Measurement
  - Effectiveness: number of matched defects / number of seeded defects.
  - Efficiency: number of matched defects per time interval (e.g., per hour)



Important: Limitations

Internal validity:

- Are observed relationships due to cause-effect releationships?
- Threats (examples):
  - Selection:
    Effect of natural variation in human performance.
    Danger: the selected group is not representative for the whole population.
  - Maturation:
    Effect of that subjects react differently as time passes.
    Examples: Subjects are being affected negatively (tired, boring) during the experiment or positively (learning effects).

External validity:

- Can findings of the study be generalized?
- Threats (examples):
  - Subjects are not representatives for population in industrial context (e.g. student experiments).
  - Objects might not be representative for industrial projects (practice).

Make study environment as realistic as possible

Example: Threats to Validity

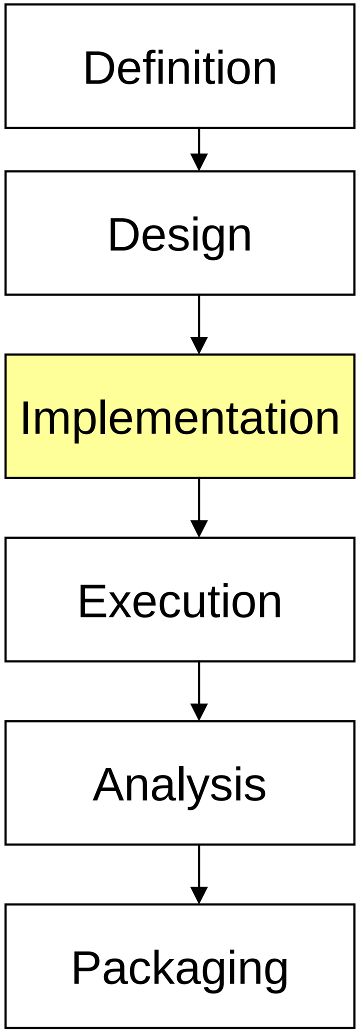
Internal Validity:

- Experience and Skills: experience questionnaire at the beginning of the experiment.
- Participant selection according to their attended course ("semi-professionals")^.
- Duration: upper time limit and allow individual (logged) breaks.
- Document package: Reviews by experts, pilot study to verify correctness.
- Etc.

External Validity:

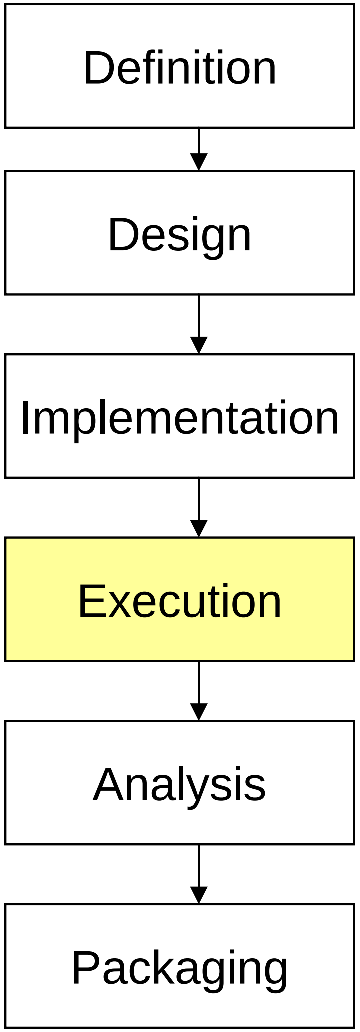
- Well-known Application domain.
- Arrangement: Classroom setting to control the experiment process.
- Participants: student experiment (might not be representative for industrial environment).
- Etc.

Experiment Process: Implementation



- Prepare material required to conduct the study.
- Use intensive reviews to check the experiment material for correctness.
- Apply Pilot-Tests to verify / improve the experiment material.
  - Are instructions clear, understandable, consistent?
  - Are tasks too simple or too difficult?
  - Can all data be collected as intended?
  - Is the schedule appropriately planned?
  - Note: participants in pilot-tests should be representative for subjects.
- Example:
  - We conducted a pilot study (including a smaller number of participants) with similar material to verify and improve the experiment package.

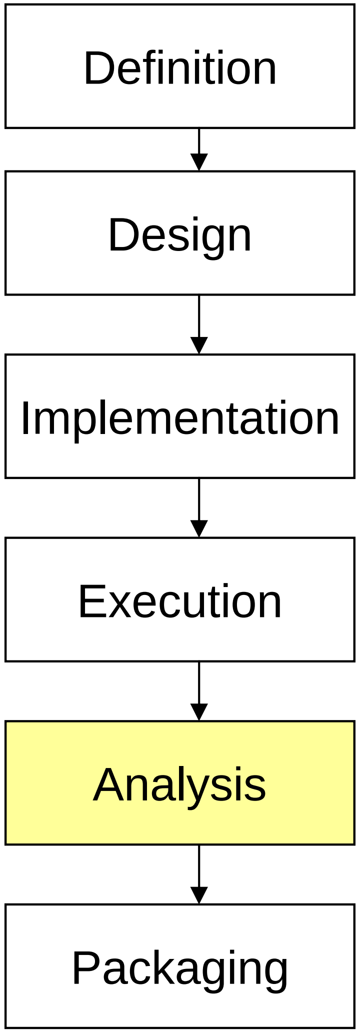
Experiment Process: Execution



- Run study according to plan and collect required data.
- Example:
  - Paper-based data collection (during the experiment)
  - Separated data submission session using a web-tool.

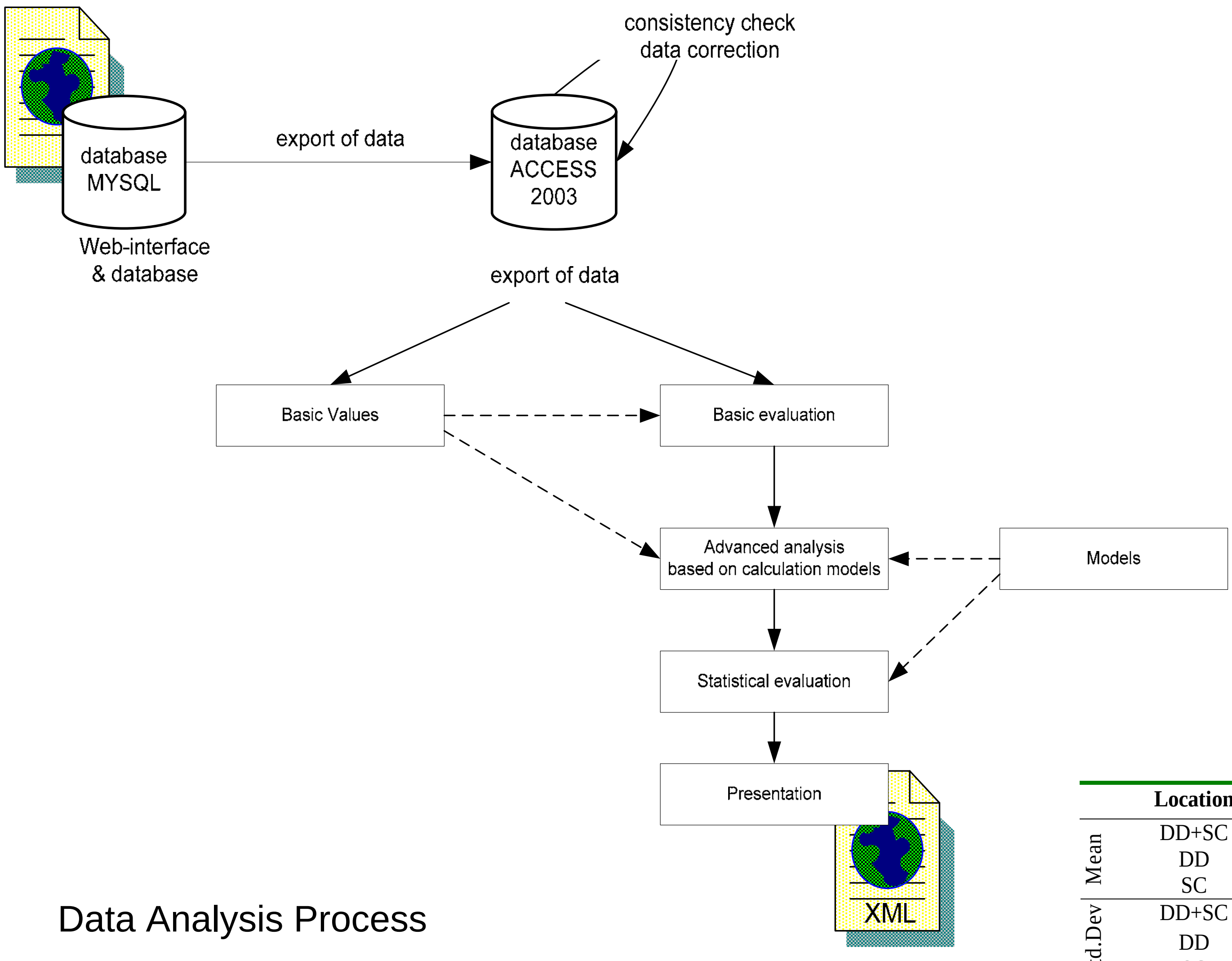
Note: this picture is from a study at ISERN 2006.

Experiment Process: Analysis

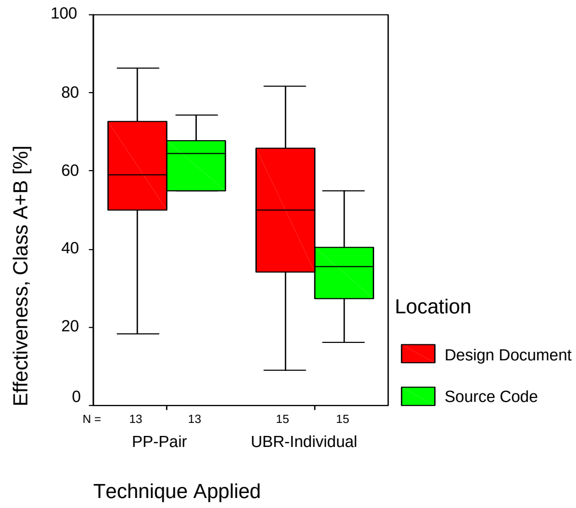


- Analyze collected data to answer operationalized study goal and hypotheses.
- Basic Steps:
  - Data collection
  - Check data for consistency and credibility
  - Create descriptive statistics and visualize data
  - Perform statistical analysis / comparison
  - Interpret results.
- Data validation ensures the correctness and completeness of collected data. Consider …
  - exceptionally high/low values, Null Values
  - Missing Values, Missing Records
  - Inconsistent values

Example: Analysis



Data Analysis Process

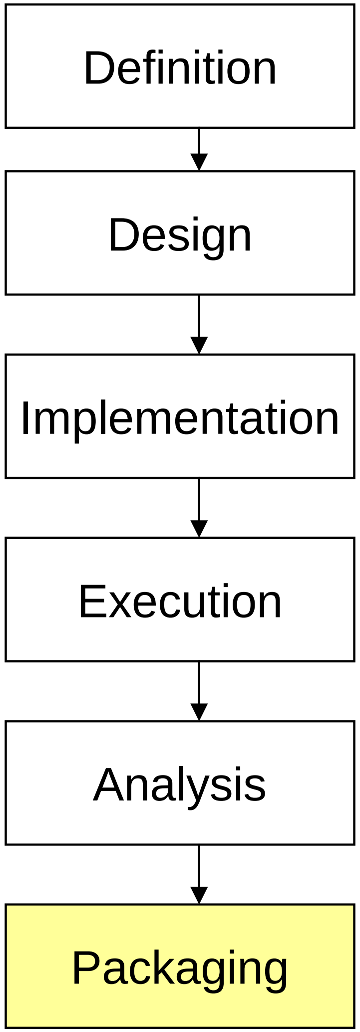


Sample Box-Plot (Pilot study)

Statistical Tests (Pilot study)

	Location	PP-Pair	UBR-Individuals	P-value
Mean	DD+SC	56.3	40.3	0.013 (S)
	DD	56.3	47.3	0.212 (-)
	SC	56.3	35.3	0.004 (S)
Std.Dev	DD+SC	20.6	13.6	-
	DD	26.7	20.6	-
	SC	17.9	11.4	-

Experiment Process: Packaging & Publication



- Report your study so that external parties are able to understand results and context of the study.
- Report your study to be replicated by others.

Sample Publications based on these study.

- S. Biffl, D. Winkler, T. Thelin, M. Höst, B. Russo, G. Succi: “Investigating the Effect of V&V and Modern Construction Techniques on Improving Software Quality”, Poster, ISERN, Los Angeles, USA, 2004.
- D. Winkler, S. Biffl: “An Empirical Study on Design Quality Improvement from Best-Practice Inspection and Pair Programming”, Profes, Amsterdam, Netherlands, 2006.
- D. Winkler, R. Varvaroi, G. Goluch, S. Biffl: “An Empirical Study on Integrating Analytical Quality Assurance into Pair Programming”, Short Paper, ISESE, Rio de Janeiro, 2006.
- D. Winkler: “Integration of Analytical Quality Assurance Methods into Agile Software Construction Practice – Research Proposal for a Family of Controlled Experiments”, IDoESE, Rio de Janeiro, Brazil, 2006.

Summary

- A study consists of a defined sequence of steps (from definition of the initial study to packaging and reporting of study results).
  - Definition: Determine study goal(s) and research hypothesis(es). Select type of empirical study to be employed.
  - Design: Operationalize study goal(s) and hypotheses.
    Make study plan: what needs to be done by whom and when.
  - Implementation: Prepare material required to conduct the study.
  - Execution: Run study according to plan and collect required data.
  - Analysis: Analyze collected data to answer operationalized study goal and hypotheses
  - Packaging: Report your study so that external parties are able to understand results and context of the study.
- A research proposal includes all relevant steps for planning, preparing, executing, analyzing, and publication of empirical studies and the results.

References

- V. Basili, G. Caldiera, D. Rombach: „The Goal Question Metric Approach“, 2000.
- S. Biffl, D. Winkler: „Value-Based Empirical Research Plan Evaluation“, Poster, ESEM, Madrid, 2007.
- B. Boehm, H.D. Rombach, M.V. Zelkowitz: „Foundations of Empirical Software Engineering – The Legacy of Victor R. Basili“, Springer, 2005.
- Freimut et al.: "Empirical Studies in Software Engineering", Tutorial, VISEK Technical Report, 2002.
- IESE Tutorials on Empirical Software Engineering.
- A. Jedlitschka and D. Pfahl: "Reporting Guidelines for Controlled Experiments in Software Engineering", ISESE, 2005.
- B. Kitchenham: “Evidence-Based Software Engineering and Systematic Literature Review”, Profes, 2006.
- C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell and A. Wesslen: "Introduction to Experimentation in Software Engineering", Kluver, 2000.
- M. V. Zelkowitz, D. R. Wallace: “Experimental Models for Validating Technology”, IEEE Computer, 1997.

Thank you for your attention

Contact:

Dipl.-Ing. Dietmar Winkler

Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstr. 9-11/188, A-1040 Vienna, Austria

dietmar.winkler@qse.ifs.tuwien.ac.at
http://qse.ifs.tuwien.ac.at

This research work has been supported by a Marie Curie Transfer of Knowledge Fellowship of the European Community's 6th Framework Programme under the contract MTKD-CT-2005-029755: CzechVMXT.