

Course "Empirical Evaluation in Informatics" Course summary, final remarks

Prof. Dr. Lutz Prechelt Freie Universität Berlin, Institut für Informatik http://www.inf.fu-berlin.de/inst/ag-se/

- Role of empiricism
- Generic method
- Concrete methods:
 - Benchmarking
 - Controlled experiment
 - Quasi-experiment
 - Survey
 - Case study

- Data analysis
- Presenting the results
- Checking study quality
- What we have not discussed



"Empirische Bewertung in der Informatik" Zusammenfassung, Nachbemerkungen

Prof. Dr. Lutz Prechelt Freie Universität Berlin, Institut für Informatik http://www.inf.fu-berlin.de/inst/ag-se/

- Rolle von Empirie
- Allgemeines Vorgehen
- Konkrete Methoden:
 - Benchmarking
 - Kontrolliertes Experiment
 - Quasi-Experiment
 - Umfrage
 - Fallstudie

- Datenanalyse
- Präsentation
- Qualitätsprüfung
- Was wir nicht besprochen haben



- Activities in Informatics research or development occur in either of three modes:
 - T Theory: Creating formal systems involving terminology and rules
 - C Construction: Creating software systems
 - E Empiricism: Learning about the characteristics of systems or processes
- Empirical methods can be applied to support both T and C
- Empirical methods can be applied to both the inputs and the outputs of T or C:
 - validating or quantifying assumptions
 - evaluating results



Roles of empiricism:

- Validation: Empirically testing whether a theory is correct
 - by testing a hypothesis derived from the theory
- **Quantification**: Empirically providing quantitative information about phenomena that are qualitatively known
 - by measuring phenomena
- Exploration: Providing insight in order to create ideas for forming theories or constructing systems
 - e.g. exploratory case studies, usability studies, benchmarking
- **Refinement**: Adding detail to theories or models
 - a form of exploration; e.g. via surveys and case studies



- In order to have impact, empirical work must be taken seriously by its audience
- In order to take it seriously, the audience must believe in its usefulness
- Usefulness is bounded by two phenomena:
 - Internal validity: The results are correct as they stand
 - External validity: The results are applicable to other contexts than those observed
- The audience's willingness to believe in validity (and thus take the work seriously) is described by:
 - **Credibility**: A reasonable degree of internal and external validty is reasonably obvious; the conclusions are warranted
 - **Relevance**: The results appear applicable to some situations of real interest

Lutz Prechelt, prechelt@inf.fu-berlin.de

Generic method

How to conduct an empirical study:

- 1. Decide on ultimate goal
- 2. Formulate question for the study
- 3. Characterize the observations sought
- 4. Design the study
- 5. Find or create the observation context
- 6. Observe
- 7. Analyze observations
- 8. Interpret results
- Credibility can be ruined in any step
- Relevance can be ruined in any of 1-5







- Each step is difficult and has potential for disaster
 - but mistakes in later steps are more easily repaired
 - so make sure you have a really good question
- Just like for software development, a waterfall model is often not a good approach to performing a study
 - Iterate all the design phases if you possibly can
 - Prototyping is almost always helpful to get a really good study
- Attacking the same question with more than one concrete empirical method increases the chances of meaningful and credible results.

Example: Deciding for/against using method X:

- Check X by a small-scale controlled experiment first,
- then perform a medium-scale case study ("pilot project")
- plus a broad-scale survey

"multi-method research"



What we have talked about in some depth:

- Benchmarking
- Controlled experiment
- Quasi-experiment
- Survey
- Case study

What we have talked about shortly:

- Simulation studies
- Literature studies
- Analysis of legacy data ("software archeology")



Criteria and illustrative *negative* examples:

- Practical feasibility
 - e.g. a controlled experiment comparing risk management methods
- Size of potential contribution to research goal
 - e.g. a survey on issues of the subconscious
- Potential for answering a relevant question successfully
 - e.g. a quasi-experiment on the project-level impact of improved compiler error messages
- Expected cost/benefit ratio
 - e.g. expensive experiments that do not generalize well
- The empiricist's skill with the method
 - e.g. doing a case study without ever having practiced qualitative research



- Description: Measure performance of a system (or method) in a standardized way; collect many results
- Advantages:
 - Objective and repeatable; high credibility
 - Results easy to understand
 - Supports accumulation of results over many studies
- Disadvantages:
 - Not practically feasible in many fields
 - Requires a shared previous understanding of the performance criteria
 - Obtaining high relevance requires much work



- Description: Change one thing, keep everything else constant, observe what happens
- Advantages:
 - The only method for proving causal relationships
 - High credibility (if done well)
 - Supports strong quantitative statistical analysis
 - Results easy to interpret
- Disadvantages:
 - Usually rather costly
 - Does not scale to large-scale, human-related questions
 - Difficult to generalize, hence relevance is dubious



- Description: Change one thing, keep everything else as constant as possible, observe what happens
- Advantages:
 - Can have very good cost/benefit ratio
 - Reasonably high credibility (if done well)
- Disadvantages:
 - Opportunistic model
 - Can often not be designed as necessary
 - Can be difficult to argue why credibility is good
 - Often difficult to generalize, hence relevance is often dubious



- Description: Ask many people what you want to know
- Advantages:
 - Cheap
 - Very flexible method
- Disadvantages:
 - Subjective; validating correctness of answers is difficult
 - Hard-to-resolve credibility problems (at least for many kinds of questions)
 - Results are almost always ambiguous

Case study



- Description: Observe something specific as it happens and broadly include as many information sources as possible
- Advantages:
 - Very rich results
 - Extremely credible (if done well)
- Disadvantages:
 - Difficult method, requires many skills
 - Generalizing any results is difficult; hence relevance is often hard to judge



- Description: Create and run an executable model of something; tweak parameters; observe
- Advantages:
 - Can investigate questions that are otherwise unfeasible
 - Flexible, cheap, yet credible and relevant (if done well)
- Disadvantages:
 - Difficult to validate correctness of the model



- Description: Review and analyze the data and/or results of several published studies together
 - In particular: Combined statistical analysis of multiple experiments ("meta analysis")
- Advantages:
 - May obtain results not possible with any one study
 - May have high robustness, hence good credibility
- Disadvantages:
 - Limitations of the given reports can not be overcome
 - Biased by non-publication of "uninteresting" results

Studies of legacy data ("software archeology")



- Description: Analyze data gathered in some pre-existing process
- Advantages:
 - Perhaps large amount of data
 - Low cost
- Disadvantages:
 - Limitations of the data can not be overcome
 - Data may be biased in difficult-to-detect ways

Data analysis



- Process of turning raw data (as collected) into results data that directly allows drawing conclusions
 - by exploring
 - by measuring
 - by comparing
 - by modeling
- Data analysis steps:
 - Make data available
 - Collect, collate, reformat, pre-process, read
 - Validate data
 - Find and correct gaps, mistakes, and inconcistencies
 - Explore data
 - Check for expected and unexpected coarse characteristics
 - Perform analysis: measure, model, or compare



Be very sceptical:

- Have some redundancy in your data
- Check redundancy
 - e.g. invariants, impossible combinations etc.
- Double-check manually entered data
- Check expectations
 - e.g. counts, frequencies, ranges, limits, etc.
- Mistrust unexpected regularities
- Mistrust unexpected irregularities
- Mistrust outliers
- Mistrust data anywhere near where you found an error



Use your common sense as much as possible!

- Make sure you understand what your variables really mean
- Formulate your expectation <u>before</u> you look at the data
- Graphics! Graphics! Graphics!
- Try out many things
- Explain to outsiders what the data are
- Ask outsiders what they think the data mean
- Ask outsiders for ideas what to analyze



- Stick to techniques you understand
 - Make sure you know (and respect) the assumptions of the techniques you use
 - If you need to think hard about what the result would mean, this is not an appropriate analysis
 - Graphics! Graphics! Graphics!
- Credibility is much more important than precision
- Validity is much more important than precision
- Illustrativeness is much more important than precision
- Get professional help if you can



Checklist:

- Do all your conclusions really contribute to answering the research question?
 - If not, are they really worth mentioning?
 - At least separate them from the others ("Further results")
- Are all your conclusions solidly backed up by your data?
 - If not, formulate them very weakly
- Can all conclusions easily be traced backwards through the study:
 - back to the analysis results
 - from there to the analysis technique
 - from there to the raw data
 - from there to the study design and setup?

Conclusion-drawing advice (2)



- Do you really trust all your conclusions?
 - If not, why should anybody else?
- Can you characterize to where you presume your conclusions generalize?
 - And why you think so? (plausibility, evidence)



A good writeup (article, technical report) or interactive presentation of an empirical study

- ...makes the elements of the generic method clearly visible
 - 1. Decide on ultimate goal
 - 2. Formulate question for the study
 - 3. Characterize the observations sought
 - 4. Design the study
 - 5. Find or create the observation context
 - 6. Observe
 - 7. Analyze observations
 - 8. Interpret results
- ...provides much detail about setup and raw data
 - perhaps in appendices



- ...uses plain, simple language wherever possible
- ...presents the data analysis in an easy-to-grasp manner
 - using graphical presentation whenever appropriate
- ...openly discusses strengths and weaknesses of the study
 - threats to internal validity
 - threats to external validity
- ...lists newly found open questions
- ...summarizes the results in the Abstract
 - rather than just announcing them

Advice:

- Prepare a rather long, detailed, comprehensive report first
- then a short version, focused on the most interesting parts.



- The quality of an empirical study is determined by its credibility and its relevance
 - If they are high, many deficiencies can be tolerated
 - If they are low, technical perfection does not help
- Good studies (which are not too common) can often be recognized quickly by these simple checks
 - Is there a clear research question at the beginning?
 - Is there a clear study result at the end?
 - Note this does not mean a clear answer.
 A good study may well be inconclusive.
 - Can the result easily be traced back to the data analysis result(s)?
 - Is the connection from analysis results to conclusion convincing?

Most bad studies are clearly bad in at least one of these aspects



What we have not (or almost not) talked about:

- Plenty of methodological details of the individual methods, e.g.
 - experiments: design of experiments
 - when manipulating more than one variable
 - surveys: instrument development
 - case studies: annotating and analyzing qualitative data
- Practical technical issues of the methods, e.g.
 - measurement infrastructure
 - calibration and validation of measurements
 - data handling and archiving
 - ethical considerations

 (e.g. privacy, copyright, informed consent)
- ...and more



Where will you apply empirical methods?

Note:

- Most diploma/master's theses can benefit a lot from a good empirical evaluation
- In fact, most would be essentially worthless without one
 - In fact, many <u>are</u> worthless for that reason
 - And many of these would never even have been done, had an evaluation been planned from the start
- Please consider this when you choose a thesis topic



Thank you!