Assessing a Version Control Study as Empirical Software Engineering Research

Thomas LaToza
What Makes Good Research in Software Engineering?
4/5/2005

Outline

• Preliminary Guidelines for Empirical Research in Software Engineering
  – Normative model of empirical research
  – What are common problems in empirical software engineering studies?
  – What should they do instead?
• Using Version Control Data to Evaluate the Impact of Software Tools….
  – Descriptive model of empirical research
  – How well does paper follow guidelines?

Preliminary Guidelines for Empirical Research in Software Engineering

• “The standard of empirical SE research is poor”
  – Identify common mistakes
  – Guidelines to make it better
  – Inspired by medical study problems & guidelines
• Authors
  – 4 SE researchers w/ “backgrounds in statistics”
  – 3 statisticians
• Statisticians defining what constitutes sound experimentation in software engineering…..

Parts of an Empirical Study

• Experimental context
• Experimental design
• Conduct of the experiment and data collection
• Analysis
• Presentation of Results
• Interpretation of Results
**What is an experiment?**

- Manipulate independent variables to understand effects on dependent variables while controlling all other variables
- **Question**
  - “Does domain knowledge help in modifying OO code more than procedural code?”
- **Hypothesis**
  - How does manipulating dependent variable affect independent variable
  - Domain knowledge helps OO developers more than procedural developers
- **Independent variables** — code: OO or procedural; domain knowledge — yes or no
- **Dependent variables** — task time, task bugs
- **Validity** — are my conclusions correct?
  - **Internal validity** — any confounds?
    - Did anything else change between conditions other than code style and the presence of domain knowledge?
  - **Construct validity** — what did I really measure?
    - Does giving them information about what the code does something about the domain?
  - **External validity** — does it matter?
    - Can the results be generalized from UIUC CS students to all CS students, programs of all sizes, all developers?

**Correlation vs. Causation**

- **Correlational study** — variable happens to be at different levels
  - Ice cream sales positively correlated to murders
  - Why?
- **Experiment** — experimentally manipulate variable
  - Give some people ice cream, others no ice cream, measure # of murders…..
- **Correlation != causality**
  - Only experiments demonstrate causality
  - “A model used to control the development process should be based on causal relationships”

**Define objectives and report information sufficient for replication**

- **Observational studies**
  - Identify factors affecting generality and applicability — applicability only to extent of similarity
    - Industry — banking, consumer goods, telecommunications, ...
    - Organization — in-house, shrink wrapped, ...
    - Skills — language, tools, domain, ...
    - Process — CMM, quality assurance, version control, ...
  - Give operationalizations of constructs
    - Everyone has slightly different definitions
    - Not just “LOC” but what counts as a LOC

- **Formal experiments**
  - Don’t oversimplify and lose context
    - Inspection effectiveness studies measured benefit in defect detection and concluded not worthwhile
    - Other studies found helpful for teamwork, knowledge transfer, training
  - Clearly state hypothesis based on an explanation
    - Shallow — “There is no correlation between cyclomatic complexity and faults found”
    - Can’t contribute to knowledge without underlying theoretical cause / effect relationship
    - Theory driven — how engineers introduce and recognize defects
  - Clearly define objectives to avoid spurious conclusions
    - Prevents having too many tests to reach statistical significance by chance
    - Prevents fishing for results in all combinations and subsets of data
### Identify the population and method of assigning treatments

- Identify how subjects / objects included
  - Random sampling necessary to prevent selection bias and generalize from sample to population
  - If not random sampling, need to justify sampling methods do not result in a sample with biased characteristics
- Random assignment for unbiased assignment of treatments
- Experimental units are independent
  - Not respondents in organization under study
- Don’t use a control unless standard baseline
  - Current industrial practice is not highly variable, not standard
  - Comparing tool against no tool is not valid comparison (not standard baseline)
- Define treatments sufficient for replicability
- Justify relationship of outcome measures to objectives
  - Changed lines of code not measurement of maintenance effort
  - Modification task time is

### Report results sufficient for replicability

- Subjective measures require interrater agreement
  - Do both raters independently make same subjective rating?
- Surveys need to report response rates and reasons for nonresponse
  - Do nonrespondents introduce systematic bias by differing from respondents?
  - Example: Cost model study respondents all completed project

### Give both characterization of data and interpretation in terms of hypothesis

- Analysis
  - Statistical tests have assumptions that must hold
  - Look at data to identify general trends and outliers which might skew statistics
- Present descriptive statistics and characterization of data
  - Results may be statistically significant, but still unimportant – Example: .1 correlation
  - Specify limitations in validity

### Application of Guidelines to a Version Control Study

- Experimental context
- Experimental design
- Conduct of the experiment and data collection
- Analysis
- Presentation of Results
- Interpretation of Results
Application to a Version Control Study

Using Version Control Data to Evaluate the Impact of Software Tools: A Case Study of the Version Editor
David L. Atkins, Thomas Ball, Todd L. Graves, and Audris Mockus

- Introduced tool that dealt with versioning in editor
- Evaluated effectiveness by looking time to complete MR for developers that did and did not use tool
  - VE tool leaves signature in checkins to detect usage

Define objectives and report information sufficient for replication

- Formal experiments -- actually longitudinal study
  - Clearly define objectives to avoid spurious conclusions
    - Does Version Editor tool reduce the effort developers spend modifying code?
  - Don’t oversimplify and lose context
    - Abstract claim “developers were 40% more productive” adding on second page “(when changing files containing preprocessor directives)"
  - Clearly state hypothesis based on an explanation
    - “Our primary hypothesis is that the VE tool reduces the effort needed to make changes involving preprocessor directives. Our secondary hypothesis is that the usage of VE would lead to shorter development intervals.”
    - No explanation!
    - But provide “anecdotal” reports of what developers found
      - Just about explanation....

Identify the population and method of assigning treatments

- Identify how subjects / objects included
  - Use all developers on SESS system for general conclusions
- Random assignment for unbiased assignment of treatments
  - Not assigning, just measuring usage of Version Editor tool
  - Acknowledge they are a correlational study
- Experimental units are independent
  - Developers are assumed to be independent; slight confound of could talk to each other and influence effectiveness of tool usage
- Don’t use a control unless standard baseline
  - Control is not using Version Editor tool
  - But not manipulating it anyway, so ok? -- guidelines unclear
- Justify relationship of outcome measures to objectives
  - Analyze at length the relationship of MR times to actual development time

Report results sufficient for replicability

- Subjective measures require interrater agreement
  - No subjective measures reported
- Surveys need to report response rates and reasons for nonresponse
  - No surveys
Give both characterization of data and interpretation in terms of hypothesis

• Analysis
  – Don’t report on specific applicability of all of their analyses – e.g. using EM in these circumstances
  – But probably ok since 2 authors are statisticians…
  – Some analyses specifically target developers with particular similar characteristics to make better comparisons

• Present descriptive statistics and characterization of data
  – Argue that it is important because of 40% number and estimate savings at 1,400 person years vs. <10 years development cost
  – Addresses validity limitations of technique at multiple points