Evaluating Your ASE Research

Jamie Andrews

Department of Computer Science
University of Western Ontario
London, Ontario, Canada
N6A 2T2
andrews@csd.uwo.ca

Plan of Talk

- The Drive for Evaluation
- Experiments
- Case Studies

Plan of Talk

- **The Drive for Evaluation**
- Experiments
- Case Studies

You and Your Research

- You have done some work on automating software engineering tasks
- You think it's pretty good
- How to convince others?
- An older way:
  - Show a small example
  - Contrast your methods with others
  - Make a persuasive argument

Problems with the Older Way

- Given techniques A and B, we can almost always find an example where A performs better than B
- How does this generalize?
- Your subjective opinion vs. referee's
- More persuasive writers win

The Drive for Evaluation

- Nowadays, good SE publication venues expect *evaluation* of research
  - TSE, TOSEM; ICSE, ASE, FSE, etc.
- Primary methods of evaluation:
  - Experiments
  - Case studies

Plan of Talk

- The Drive for Evaluation
- **Experiments**
- Case Studies

Experiments

- Characteristics of experiments
- Objective measures
- Statistical analysis
- Threats to validity
- Experiment design

Characteristics of Experiments

- Subjects: things you are changing / using / acting on / treating / helping
  - Programmers, projects, programs, specs, test suites, ...
- Treatments: things you are doing associated with your research work
  - Tools used, processes followed, program analysis techniques, test case generation techniques, ...

Classic Experiments

- Classic experiments:
  - Subjects drawn from target population
  - Subjects selected randomly
- Examples:
  - Subjects = patients;
    treatments = different drugs
  - Subjects = trees;
    treatments = different fertilizers

SE Experiments with Human Subjects

- Experimental subjects often not drawn from target population
  - Targets: programmers in industry
  - Subjects: students
- Differences between subjects and targets unpredictable

SE Experiments with Human Subjects

- Challenges:
  - Ethics approval
  - Assigning subjects to treatment groups
  - Subjects learn, change
  - Subjects drop out

SE Experiments with Non-Human Subjects

- Subjects often not randomly drawn from target population; rather chosen due to simple availability
  - Technically "pseudoexperiments"
- Main challenge: Subject preparation can be tedious, little reward

Objective Measures

- Numerical measures not based on experimenter's opinion
- Examples:
  - Likert-scale ("on a scale of 1 to 7,...") answers from subjects
  - Cost: CPU time, clock time, effort, number of test cases, size of model
  - Effectiveness: Coverage, accuracy, precision / recall, number of bugs per KLOC

Statistical Analysis

- Visualizations
- Comparisons
- Correlations

Visualizations

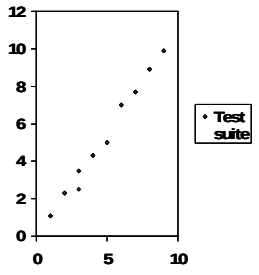
- Scatter plots, line plots, bar graphs, pie charts
  - Can be done by **`gnuplot`** or Excel
- Box plots, more complex visualizations: use a statistics package
  - SPSS, Minitab, R

Comparisons

- Comparing mean, standard deviation of data set A and data set B not enough
  - Must take into account number of data points
- Student's *t* test was the first to do this
  - Available on Excel
  - Paired vs. unpaired versions
  - If *p* value < 0.5, "statistically significant"
- Wilcoxon ("Mann-Whitney", "rank sum") also useful
  - Makes fewer assumptions about data

Correlations

- Pearson (standard) correlation
- Spearman correlation
- Kendall correlation
- "Correlation is not causation!"



Threats to Validity

- No experiment perfect
- Various possible ways that results of experiment may not generalize
  - Technical term for these ways: "threats to validity"
- Modern approach: openly discuss:
  - Threats to validity
  - Any measures taken to minimize them

Experiment Design

- An milligram of preparation is worth a kilogram of:
  - Having to re-run experiments
  - Having to justify weak results
- Books / courses on experiment design

Plan of Talk

- The Drive for Evaluation
- Experiments
- **Case Studies**

Experiments are Infeasible / Impractical When...

- Research is at an initial stage
- Result uncertain and experiment would take a long time
- Subjects are inaccessible / too large
  - e.g. product lines
- "Case study" may be an option

Examples vs. Case Studies

Examples	Case Studies
• All aspects under researcher's control	• Some aspects not under researcher's control
• Used to illustrate technique	• Used to evaluate technique in practice
• Post-hoc, ad-hoc, selective detail	• Preparation, detailed analysis of data

Elements of Case Studies

- Research questions
- Expectations
- Quantitative and qualitative data wanted
- Subject selection
- Data analysis

Example Case Study

- Research questions:
  - Was our tool useful and easy to use for software engineers?
  - In what ways did they find it helped?
  - What problems (if any) did they have with it?
- Expectations:
  - It was useful in understanding code design
  - There are some known usability problems

Example Case Study

- Quantitative and qualitative data wanted:
  - How often they needed to use the tool
  - How long it took them to do the tasks
  - Their reactions to the tool
- Subjects:
  - Two software engineers at company X, each with over 5 years of experience

Example Case Study

- Writeup includes:
  - Raw data on frequencies of use, amounts of time taken
  - Quotations, reactions from software engineers

References and Resources

- Experiments:
  - Book: Claes Wohlin et al., *Experimentation in Software Engineering: An Introduction*, Springer, 2000
  - Paper: Barbara Kitchenham et al., "Preliminary Guidelines for Empirical Research in Software Engineering", *IEEE Trans. Software Eng.,* 28(8): 721-734 (2002)
  - Example papers: most papers by those authors and by Basili, Briand, Harrold, Ostrand, Rothermel

References and Resources

- Case studies:
  - Classic text: Robert K. Yin, *Case Study Research: Design and Methods*, 3rd ed., Sage Publications, 2003
  - Example: Baniassad et al., "Design Pattern Rationale Graphs: Linking Design to Source", *ICSE 2003*

Thanks to...

- **`wpclipart.com`** for the open source images
- You, the audience
  - Questions?