Anatomy of an Experiment: Difficulties in Evaluating Rationale-based Systems

Janet Burge
Miami University
Evaluation...

• First, you have a hypothesis….  
  – Then you need evidence!
• DR systems are used by “real” (human) designers
  – Requiring empirical studies
Some background...

• Software Engineering Using RATionale (SEURAT)
  – Integrates Rationale capture and use into the Eclipse Java IDE to support software maintenance
  – Inferences over rationale to detect problems and perform impact assessment
  – Key focus is on uses for rationale
SEURAT Evaluation

• Evaluated using 20 subjects from academia and industry
• Experiments were conducted over a two month period
• Results were promising but not definitive
Experiment Design

• Goal-Question-Metric (GQM) analysis
  – What is the goal of our evaluation?
  – What questions do we need to ask to see if we have met that goal?
  – What metrics can we collect to answer the questions?
Experiment Testing

• Dry runs with two subjects at different ends of the expertise spectrum

• Results
  – Usability enhancements to the system
  – Serious scaling back of the experiment to meet time constraints
Experiment Subjects

• Difficult to recruit even with bribes
• Questionnaires used to assess expertise and experience
• Two groups: experimental and control were formed with ten subjects each
Experiment Administration

• Observed tests
• Introductions to Eclipse
• Tutorial on SEURAT to experimental group (using rationale for a different system)
• Three maintenance tests in a random order
• All subjects were allowed to complete (no matter how long it took....)
Results

• Some things couldn’t be tested
  – Like how long it took to use SEURAT to find problems

• SEURAT improved performance of those with some or moderate Java experience (not experts!)

• No statistical significance…
Issues

• Task Problems
  – Small tasks increase noise problems
  – Some features couldn’t be tested using a control group

• Subject Availability
  – Not enough subjects for significance…

• Time Problems
  – Even simple tasks were time consuming
  – Large variances in control group (control “some” group: 6.7, 32, 52 minutes for one task)

• Noise Problems
  – Learning time a large percentage of total time
  – Experts wanting to “play” with SEURAT first
  – Time required to update rationale in addition to code
How to Fix?

• Real project, experienced developers
  – But how do you collect the initial rationale base?
  – Wouldn’t collection participation bias subjects?
  – How can you compare two groups on the same task? (waste of company time)
  – Timing data over the long term?
Conclusions

• Small, laboratory experiments are still needed
  – Multiple experiments with same subjects to cancel learning effects
  – More experienced subjects and more time would allow more difficult tasks
  – Larger number of subjects – more likely to find significant results and outliers can be eliminated

• Industrial experiments are still needed too
  – Perhaps as a more qualitative measure?