EMPIRICAL RESEARCH METHODS IN VISUALIZATION

... and some thoughts on their role in Master, PHD and postdoctoral projects

Talk at University of Sydney, 11. August 2014
My Background

The Java Code Clone Detection API
Evaluation of Visualization Techniques

Typical question:
Is a given visualization technique appropriate for a given application scenario?

- Application Scenario
  - Data: What kind of data is to be visualized?
  - Task: What task does the user want to solve?
Quantitative/qualitative Evaluation

**Quantitative**
- Analysis of numerical data
- Goal: exact description of the user behavior
- Data acquisition:
  - Tests, Tasks
  - Questionnaire

**Qualitative**
- Analysis of textual, audio-visual data
- Goal: Understanding of the user behavior
- Data acquisition:
  - Observing the user
  - Interviews, open-ended questions
Evaluation: When and why?

Explorative Study
• Goal: elicit requirements, generate hypotheses
• qualitative/quantitative

Formative Study
• Goal: improve visualization technique during development
• usually qualitative

Summative Study
• Goal: validate hypotheses
• usually quantitative

development process
# Evaluation: Criteria

What does „better“ or „more appropriate“ mean?

<table>
<thead>
<tr>
<th>Effectivity</th>
<th>Efficiency</th>
<th>Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Was the task solved?</td>
<td>• How long did it take to solve the task?</td>
<td>• Subjective assessment/experience of the user (questionnaire)</td>
</tr>
<tr>
<td>• Was new knowledge/skill learned? (Pre-Test/Post-Test)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Qualitative Evaluation
Prototyping

- Kinds of Prototypes
  - Verbal description
  - Pen-and-Paper
  - Partial implementation
    - c.f. vertical/horizontal prototypes in software engineering
Guideline Checking

- Walk through a list of design recommendations
- No test persons required
- Easy to do

Guidelines
- Usability Guidelines
- Focus+Context
- InfoVis Mantra
- Gestalt Laws
- Aesthetic Criteria (graph drawing)
- ...

Cognitive-Dimensions Framework

- Identify generic tasks to be solved with the tool/technique
- Assess for each task how relevant each cognitive dimension is.
- Compare to ideal profile
- No test person required

Cognitive Dimensions

- **Viscosity** = Resistance to change
- **Closeness of mapping** = Closeness of representation to domain
- **Premature commitment** = Constraints on the order of doing things
- **Hidden dependencies** = Important links between entities are not visible

Other Cognitive Dimension are **Visibility**, **Role expressiveness**, **Error-proneness**, **Abstraction**, **Secondary notation**, **Consistency**, **Diffuseness**, **Hard mental operations**, **Provisionality**, and **Progressive evaluation**

http://www.cl.cam.ac.uk/~afb21/CognitiveDimensions/
Thinking Aloud

- Participants are asked to verbalize their thoughts aloud
  - Goals
  - Questions, Problems
  - Decisions
- Audio (and Video) recording
- Small number of test persons

http://www.measuringusability.com/blog/thinking-aloud-time.php
Grounded Theory

- Develop a „theory“ based on qualitative data (grounded on observations)
- Iterative process of gathering and analysing data
- Stepwise refinement of a theory (see diagram)
- Small number of test persons
Example: Graph Merging

How do users merge two similar UML class diagrams?

[Rutz et al., 2011]

Figure 3. Relations between top level categories. Continuous arrows encode preconditions and dashed arrows depict supportive relations.
Quantitative Evaluation
Controlled Experiment

- Controlled experiment (laboratory conditions)
- Variables
  - **independent variables**: controlled properties (systematically changed)
  - **dependent variables**: measured properties
  - **co-variables**: additional, non-controlled factors, which may have an impact on the dependent variables
Experimental Design

- Treatment
  - one variation of the independent variables

- Between Subject
  - One group per experimental condition
  - (Informed) Randomization to balance groups

- Within Subject
  - Every test person performs all experimental conditions
  - Randomization to prevent training and fatigue effects
Hypotheses

- **Null hypothesis**: There is **no** effect
  - **no** impact of the independent variables on the dependent variables

- **Alternative hypothesis**: There is **an** effect
  - Impact of independent variables on the dependent variables

- **Goal of Experiment**: Rejecting the null hypothesis
Statistical Analysis

- Descriptive statistics: mean, median, quartiles, etc.
- Inference statistics
  - Statistical tests to validate hypotheses
  - Assume alternative hypothesis if null hypothesis can be rejected with sufficient safety (statistically significant effect)
  - Significance level: error probability of assuming the alternative hypothesis although the null hypothesis is valid (alpha-error)
- Experimental design, effect size, and significance level determine the minimum number of test persons
Criteria for selecting a test

- type (categorial/numerical) and number of dependent variables
- Experimental design (within/between subject)
- parametric/non-parametric: normal distribution

Tests

- Wilcoxon: non-parametric test of two related treatments (within subject)
- t-Test: parametric alternative to Wilcoxon (both within and between subject)
- ANOVA: to be used instead of t-Test for more than two treatments

Decision tree by Field and Hole, 2002
Example: Radial vs. Cartesian

Can users better remember information presented in radial or cartesian visualizations?

- quantitative
- „within subject“
- summative: test of hypotheses derived from theories
- explorative: generation of new hypotheses (observations)
- Independent variables: radial/non-radial, task, size, background pattern
- Dependent variables: correctness and response time
- Wilcoxon tests and ANOVA

[Diehl et al., 2010]
Summary

Goal of Evaluation
- explorative
- formative
- summative

Qualitative Evaluation
- Textual or audio-visual Data
- Understanding
- Multitude of methods

Quantitative Evaluation
- Numerical data
- Description
- Mostly controlled experiment
How long does it take?

- Grounded Theory Study „Merging UML Diagrams“
  - three people (1 bachelor student, 1 PhD student, me)
  - more than half a year (most of it for analysis)

- Web-based Controlled Experiment: „Radial vs. Cartesian“
  - Three people (2 PhD students, me)
  - Several weeks for experiment, several weeks for analysis

- In comparison: Our work on mining software repositories to generate recommendations for programmers [Zimmermann et al., 2005]:
  - No test persons, used historical data as an oracle, evaluation took several days (to write the scripts for the evaluation)
My five cents of advice

- Bachelor or master thesis
  - Light-weight qualitative approaches
  - Quantitative: experimental design and a pre-study

- PhD thesis
  - Qualitative and quantitative evaluation
  - But beware: you may fail to get interesting results and thus the results may be difficult to publish, so do not put an empirical study in the center of your thesis.

- Post Doc
  - Qualitative and quantitative evaluation
  - Can take the risk of failure
References

[Diehl et al., 2010] Uncovering Strengths and Weaknesses of Radial Visualizations -- an Empirical Approach
Stephan Diehl, Fabian Beck and Michael Burch
IEEE Information Visualization Conference 2010, Salt Lake City, Utah, USA, October 2010.

[Lutz et al., 2011] How Humans merge UML-Models
Rainer Lutz, David Würfel and Stephan Diehl

[Zimmermann et al., 2005] Mining Version Histories to Guide Software Changes
Thomas Zimmermann, Peter Weißgerber, Stephan Diehl and Andreas Zeller. In IEEE Transactions on Software Engineering, 31(6), June 2005. (received ICSE 2014 Most Influential Paper Award)