Handout for USER Workshop (User Evaluation for Software Engineering Researchers), ICSE, Zürich, Switzerland, June 2012

The USER organizers and expert panel

Introductory Note

A list of the most important properties of some important methods. For rough orientation only, please consult detailed instructional material before you apply a method.
The description of each method will have the following parts:

Definition: What are the defining properties of the method?
Scope: When does the method apply?
Strengths: The nice properties of executing the method or of the results you can obtain
Limitations: Where or when the method does not work well; what to use instead
Issues: Where or when the method is difficult to use; what to do about this
Pitfalls: Typical mistakes to avoid

"Complete" methods

These methods link certain types of data (and perhaps modes of their collection) to certain modes of data analysis and certain types of result to be obtained. They employ the data collection methods and data analysis methods described further down.

Controlled Statistical Experiment

Definition: Measuring the causal influence of an independent variable by keeping all other independent variables (quasi) constant via repetition, randomized assignment, and averaging (together these constitute the "control").
Scope: Use this when all of the following are true: (1) you have a binary question to answer (e.g., is tool x better than tool y?), (2) have the ability to control the environment (e.g., have access to a lab in which your participants can conduct the task), and (3) you can gather enough data to support inferential statistical analyses. A rough rule of thumb is 30 participants for each treatment.
Strengths: Reviewers tend to value these the most highly.
Limitations: Not suitable for answering "how" or details-oriented "why" sorts of questions.
Issues: Must have enough participants to enable statistically significant results.
Pitfalls: Confounding effects that invalidate the results (For example, the learning effect means that participants may have an easier time using the second system in a study because it is second, not because of the system).

Quasi-Experiment

Definition: Like controlled statistical experiment, except that constancy is incomplete, typically due to lack of randomization.
Scope: Applies instead of a controlled statistical experiment when the independent variable/treatment cannot be assigned by you. For example, you cannot assign gender, so gender studies fall in this category. Medical studies are almost all of this type, because they cannot assign variables like "has had cancer".
Strengths: Same as controlled statistical experiments.

Limitations: Same as controlled statistical experiments.
Issues: Same as controlled statistical experiments.
Pitfalls: Same as controlled statistical experiments.

Controlled Qualitative Study

Definition: Controlled study, but no statistics. (Therefore it would be technically incorrect to call it an "experiment").
Scope: Use this kind of study when you want qualitative (non-numeric) information, such as lists of phenomena or answers to "how" or "why" questions. (e.g., what barriers do users face when they try to debug?). Produces qualitative data such as lists of behaviors, verbal comments, etc. Usually done with only a handful of participants, since each one produces a great deal of data.
Strengths: These studies reveal detailed problems and surprising phenomena that were not predicted in advance, and as a result often produce prescriptive design implications.
Limitations: Not suitable for answering binary questions (eg, is x better than y?).
Issues: Qualitative analysis is very time-consuming.
Pitfalls: Lack of rigor in the analysis procedures.

Survey

Definition: Collecting information from many respondents using a questionnaire
Scope: Gather quantitative frequency data for demographic or memorable information and examine frequencies and correlations. Gather large amounts of qualitative data from open-ended questions.
Strengths: Usually lightweight. Can gather lots of information quickly. Useful for seeing if trends from small samples (e.g., observations, interviews) generalize to a larger population. Also useful for identifying low-frequency, salient events.
Limitations: Not suitable for eliciting information about participants they cannot recall (e.g., non-salient events). Experimenter interaction with respondents is one way - often no opportunity to follow-up on ambiguous or interesting responses to questions, unlike observations or interviews.
Issues: Responses biased by many factors - e.g., question wording, non-response. Reported events likely to be biased towards salient events.
Pitfalls: Must be careful to recruit participants that generalize to desired population. Should carefully pilot questionnaire to ensure it measures what you expect.

Case Study

Definition: Forming or evaluating multiple competing explanations of a phenomenon by using multiple sources of evidence and employing triangulation.
Scope: Answering "how" or "why" questions about a complex phenomenon as it unfolds in its natural context.
Strengths: Evaluates the research in-context in a realistic setting.
Limitations: May not be the best way to compare different techniques/tools.
Issues: Results may not generalize, especially if the number of participants is small.

Pitfalls: Using a tool/technique/experiment designer as the subject of the case study -- this produces biased results and is not a true case study. Many "case studies" appearing in publication are illustrative examples rather than actual case studies.

Ethnography

Definition: Describe a social system based on extensive participant observation
Scope: Sensemaking of cultures (complex social processes)
Strengths: Extremely detailed understanding of the factors that govern behavior, useful for understanding how a tool might be used in a particular culture.
Limitations: Not good for answering comparative questions
Issues: Extremely time consuming to perform (often months or years). Some value can be extracted from brief observations, but these often give only a limited glimpse of the culture of a place.
Pitfalls: These are not interviews, nor are they interventions; the ethnographer's job is to observe and probe.

Cognitive Modeling

Definition: Formulate models and predictions of user behavior based on cognitive theories
Scope: Predicting user behavior and performance
Strengths: Theory-driven hypotheses are easily explainable and generate new predictions.
Limitations: Models may not account for deviations or outliers.
Issues: Many theories are not yet well-validated or tested in software engineering domain.
Pitfalls: Broad strokes painted by cognitive theories may miss finer motivations and circumstances.

Participatory Action Research

Definition: Focus on the effects on how the researcher's direct actions participating within the community of study can achieve the goal of improving the community's performance or quality of an area of concern.
Scope: When the community cannot act alone adequately in this matter
Strengths: Innovation
Limitations: Generalization of results is generally difficult
Issues: Keep the roles of researcher and community clearly apart
Pitfalls: Not being able to describe properly what was done or how/why it was research

Data collection methods

Questionnaire

Definition: Ask many participants a predetermined series of questions ("items"), often through a web

form.

Scope: Quantitative data through forced choice items. Qualitative data through free response items. Quantitative data through content analysis of free response items.

Strengths: Can probe many (hundreds or even thousands) of participants with no extra data collection effort. Removes all bias from direct interactions with the experimenter.

Limitations: Interaction with respondents is one way - usually cannot follow up on responses.

Issues: Can examine correlations between measures without any sense of why they occurred or what they mean.

Pitfalls: Need to use appropriate statistical tests when examining relationships between variables. Mean is not a meaningful measure for skewed distributions. Outliers can skew results. Participants that incorrectly interpret items can lead to bogus results.

Direct Observation

Definition: Observe events personally as they occur.

Scope: All qualitative modes of inquiry and low-bandwidth quantitative ones.

Strengths: Rich information is available, unexpected events can be recognized.

Limitations: Some environments may not allow data recording which puts the burden of accuracy on the researcher.

Issues: Researcher is overwhelmed quickly.

Pitfalls: Presence of researcher disrupts normal peer interaction (oh you're busy with someone). Hawthorne effect - knowing that you are being observed can change the results.

Instrumentation, Automated Measurement

Definition: Record events in programming environment

Scope: Need empirical measures of user behavior and performance

Strengths: Generates definitive quantitative evidence of phenomena.

Limitations: Limited insight in interpreting events.

Issues: Scalability; Privacy concerns; IDE instrumentation tricky

Pitfalls: Always one more thing that you wished you recorded

Audio/Video Recording

Definition: Record audio, screen capture video, or room-level video

Scope: Need record but don't know what to measure yet. Want to use content analysis for observations or interviews. Want to analyze task or subtask time in detail post-hoc.

Strengths: Playback of user study lets you take high-level notes and revisit detail later. Essential (with think aloud) for understanding participant behavior in detail to understand why something occurred.

Limitations: Limited precision in interpreting events. Unlike interpretation during observation or interview, can´t use participant to help interpret.

Issues: Time consuming to aggregate data using content analysis, open to subjective measures and interpretations. Can be greater concerns in field settings with NDAs and confidentiality due to increased fidelity of data.

Pitfalls: Often easy to capture, but time consuming to analyze in detail.

Diaries

Definition: Research subjects record actions and events as they occur.
Scope: Provides qualitative data and potentially quantitative (if multiple subjects record information consistently)
Strengths: Activities are recorded as they happen, could identify activities not anticipated by researcher. Requires less researcher time than observation.
Limitations: Time sensitive tasks where the delays introduced by recording will impact task success.
Issues: Self-reported data may not be accurate or complete.
Pitfalls: Recording is intrusive and the act of recording actions is likely to affect behavior.

Data Generation Methods

Interview (Semi-structured or Structured)

Definition: One-on-one discussion with participant
Scope: Want to find out detailed information about what participants are thinking.
Strengths: Participant more likely to open up; can get information about thought processes.
Limitations: Participants may not accurately remember what they did when asked about it later; can combine interviews with other techniques and have participants also perform tasks.
Issues: Can be difficult to ask the right questions.
Pitfalls: Perform semi-structure interview with small pool; then structured with large pool

Focus Groups

Definition: Small pool of participants give feedback on prepared material (e.g. paper prototypes)
Scope: Need to quickly get feedback or brainstorm ideas with intended users.
Strengths: Group members can bounce ideas off of each other.
Limitations: Groupthink or dominating participant derails discussion.
Issues: Moderating discussion is an art.
Pitfalls: Not enough control over the discussion, too large of a group.

Think-aloud

Definition: Participant vocalizes thought processes during study.
Scope: Need to observe the decision making processes of people.
Strengths: Explains user behavior; gives deep insight into cognitive processes during task
Limitations: Self-reported introspection may introduce bias in task; not every human thought can be verbalized (high-level rational thought is okay, perceptual processes are not).
Issues: Transcription is time consuming. Must occasionally probe participants to speak up
Pitfalls: Failure to correlate transcription events with current activity/screen

Peer-Exploration

Definition: Observing participants working together on tasks (e.g., programming tasks) as a pair rather than individually.
Scope: Useful for studies where participants' thoughts are needed to provide more color around "what" they are doing and "why".
Strengths: Provides more genuine discourse than think-aloud.
Limitations: There are limits to how well thoughts can be verbalized.
Issues: The presence of a pair can influence the behaviour or a participant in ways that affect applicability of the results. Not all tasks lend themselves to the use of pairs of participants.
Pitfalls: One participant taking over the task can reduce the amount of verbalization.

Contextual Inquiry

Definition: Researchers observe participants in their place of work, asking probing questions about their processes, decisions, and tasks.
Scope: Useful for observing how people currently work in order to inform the design of new workflows and technologies.
Strengths: Can be done in as little as a couple hours for 1 participant.
Limitations: Because of the small samples (less than 10, typically), can be difficult to generalize.
Issues: Works best when the workplace is in one place; less useful when the work is entirely on a screen, because it can be difficult to see the screen.
Pitfalls: Researchers often treat these as interviews, and participants often think they are interviews. Needs explicit instruction before hand to describe master/apprentice dynamic (the participant is the master, teaching the researcher).

Data Analysis Methods

Visualization

Definition: Represent or summarize data by appropriate graphical plots or diagrams.
Scope: For non-tiny amounts of quantitative data whenever overview and insight trump detail and precision (that is, always) or for qualitative data to tell a long story short
Strengths: Can be understood quickly; peculiarities remain visible; can tell a story
Limitations: Incongruence when comparing to other studies; limited generalizability
Issues: For ill choice of visualization type: Looks can be misleading
Pitfalls: Readers may be able to spot minor inconsistencies you have overlooked that are otherwise hidden

Statistical Evaluation

Definition: Apply statistical inference to draw conclusions from quantitative data.
Scope: All of the above
Strengths: Often considered very credible, “wow” factor; reliable insights
Limitations: Requires assumptions that may be hard to validate
Issues: May turn interesting data into what is perceived as a null result.

Pitfalls: Violated assumptions, distorted interpretations, taming variance in user skills

Discourse Analysis (Content Analysis)

Definition: Analyze qualitative data (typically verbal communication) and quantify "Who says what, to whom, why, to what extent and with what effect?" by relying on a strict *coding scheme*.
Scope: Quantification of event frequencies in qualitatively well-understood domains.
Strengths: Repeatable
Limitations: Involves simplification
Issues: Unreliable unless coding scheme is very well worked out, likely requiring iteration.
Pitfalls: Not recognizing if the coding scheme does not fit the data well; using coders that are not trained well.

Grounded Theory

Definition: Qualitative data analysis that generates (as opposed to test) theory by abduction, constant comparison, and theoretical sampling.
Scope: Useful for exploring understudied phenomenon.
Strengths: Does not require researcher to know a priori what constructs or concepts will be useful for understanding the situation being studied. Important concepts can be "discovered" through the analytic process.
Limitations: Because of small samples, can be difficult to generalize.
Issues: Detailed open coding can be time consuming and difficult to repeat. Also, in my experience it can be difficult for new researchers to find interesting insights in the data.
Pitfalls: Getting lost in detailed coding and missing the important questions to be asked.

References

A. Dean and D. Voss. Design and Analysis of Experiments. Springer, 1998.
Good introduction to the statistical data analysis side of experiment design.

IDEO. Method Cards: 51 Ways to Inspire Design. William Stout, 2003. (available as iPhone app!)
This set of cards introduces 51 evaluation methods to try out.

B. Kitchenham. Evaluating software engineering methods and tool, Part 1: The evaluation context and evaluation methods. ACM SIGSOFT Software Engineering Notes, 21(1):11–14, 1996.
A series of articles by Barbara Kitchenham focus on applying a variety of user evaluation methods to the software engineering context.

H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Seven Guiding Scenarios for Information Visualization Evaluation. Technical Report 2011-992-04, University of Calgary, 2011.
A broad overview of evaluation techniques, grounded in the context of information visualization. Organized around a variety of evaluation scenarios (e.g. evaluating user performance) and describes questions, methodologies, and analyses for each scenario.

J. Lazar, J. H. Feng, and H. Hochheiser. Research Methods in Human Computer Interaction. Wiley, 2010.
This book contains a good overview of various evaluation methodologies (such as surveys, diaries, interviews, ethnography, etc). For each methodology, describes issues, design choices, research questions, and analyses. Many suggestions for recommended reading.

T. C. Lethbridge, S.E. Sim, J. Singer. Studying Software Engineers: Data Collection Techniques for Software Field Studies. Empirical Software Engineering, 10, 311=311, 2005. *This paper provides a taxonomy of data collection techniques for software engineering field studies. Advantages and disadvantages are presented for each technique.*

J. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In Human-computer interaction, pages 152–169. Morgan Kaufmann Publishers Inc., 1995.
This classic paper introduces a taxonomy of study strategies, and may be useful for determining where on the spectrum a study should be to answer a particular research question.

J. Nielsen. Usability Inspection Methods. Wiley, 1994.
This classic book introduces a variety of lightweight "discount" usability testing methods.

A. Oram, G. Wilson, eds. Making Software: What Really Works, and Why We Believe It. O'Reilly. 2010.
A great compilation of articles about empirical experiments in software engineering. Some of them evaluate software, some evaluate software practices, and others study software developers themselves.

F. Shull, J. Singer, and D. Sjoberg. Guide to Advanced Empirical Software Engineering. Springer 2008.
Applying empirical qualitative and quantitative methods to understanding software development practices.

D. Sjoberg, T. Dyba, and M. Jorgensen. The future of empirical methods in software engineering research. In Future of Software Engineering (FOSE), 2007.
This call-to-arms incorporates a systematic review of empirical methods for software engineering research and includes more than 130 citations.

C. Snyder. Paper Prototyping: The fast and easy way to design and refine user interfaces. Morgan Kaufmann, 2003.
This excellent book has step-by-step instructions on the design, implementation, and user testing of prototypes, using the medium of paper as an example.

The Visual Guide to Cognitive Biases.
http://www.scribd.com/doc/30548590/Cognitive-Biases-A-Visual-Study-Guide
A great overview of common biases to watch out for when designing studies.