

University of Toronto Department of Computer Science

Research Methodology Session

Steve Easterbrook
with help from:
Dana, Martin, Axel, Steve F., Robyn, ...

© 2003, Steve Easterbrook 1

University of Toronto Department of Computer Science

Motivation

- ▷ Frustration with reviewing for RE conferences
 - ↳ papers generally don't talk about evaluation of research results
- ▷ New work on Benchmarks for SE
 - ↳ see: Susan Sim's PhD thesis, UofT, April 2003.
 - ↳ process of creating benchmark makes a research community more (scientifically) mature
- ▷ Questions:
 - ↳ How do we choose our research goals?
 - ↳ How do we evaluate success?
 - ↳ How do we measure the impact/importance of a research program?
 - ↳ Should we be more explicit about our research methods?

© 2003, Steve Easterbrook 2

University of Toronto Department of Computer Science

A collection of 'idioms'

- ▷ Toy problems
- ▷ Exemplars
- ▷ Benchmarks
- ▷ Field Studies
- ▷ Action Research
- ▷ Laboratory experiments
- ▷ Rational Reconstructions
- ▷ Pilot Studies

Good for?

- Setting research goals
- Didactic illustration
- Validating results

© 2003, Steve Easterbrook 3

University of Toronto Department of Computer Science

Exemplars

"self-contained, informal descriptions of a problem in some application domain; exemplars are to be considered immutable; the specifier must do the best she can to produce a specification from the problem statement."

Good for:

- Setting research goals,
- Understanding differences between research programs

Limitations:

- No clear criteria for comparing approaches
- Not clear that "immutability" is respected in practice

Examples:

Meeting Scheduler; Library System; Elevator Control System; Telephones;...

see:

M. S. Feather, S. Fickas, A. Finkelstein, and A. van Lamsweerde, "Requirements and Specification Exemplars," Automated Software Engineering, vol. 4, pp. 419-438, 1997.

© 2003, Steve Easterbrook 4



Benchmarks

"a test or set of tests used to compare alternative tools or techniques. A benchmark comprises a motivating comparison, a task sample, and a set of performance measures"

good for

making detailed comparisons between methods/tools
increasing the (scientific) maturity of a research community
building consensus over the valid problems and approaches to them

limitations

can only be applied if the community is ready
become less useful / redundant as the research paradigm evolves

examples

TREC Ad Hoc Retrieval Task; C++ fact extraction;

See:

S. Sim, S. M. Easterbrook and R. C. Holt "Using Benchmarking to Advance Research: A Challenge to Software Engineering". Proceedings, ICSE-2003



Field Studies

exploratory study, used where little is currently known about a problem, or where we wish to check that our research goals are grounded in real-life settings; studies organisational practice using anthropological techniques.

good for

setting a research agenda (what really matters?)
understanding the context for RE problems (naturalistic inquiry)

limitations

hard to build generalizations (results may be organisation specific)
observers' bias

examples

Curtis et al; studies of globally distributed development; ...

See:

Klein, H.K., and Myers, M.D. "A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems," MIS Quarterly, vol 23 No 1, pp67-93, 1999.



Action Research

"research and practice intertwine and shape one another. The researcher mixes research and intervention and involves organizational members as participants in and shapers of the research objectives"

good for

any domain where you cannot isolate {variables, cause from effect, ...}
ensuring research goals are relevant

limitations

hard to build generalizations (abstractionism vs. contextualism)
won't satisfy the positivists!

examples

most of RE???

See:

Kock, N.F., (1997), Myths in Organisational Action Research: Reflections on a Study of Computer-Supported Process Redesign Groups, Organizations & Society, V.4, No.9, pp. 65-91.



Laboratory Experiments

experimental investigation of a testable hypothesis, in which conditions are set up to isolate the variables of interest ("independent variables") and test how they affect certain measurable outcomes (the "dependent variables")

good for

quantitative analysis of benefits of a particular tool/technique
(demonstrating how scientific we are!)

limitations

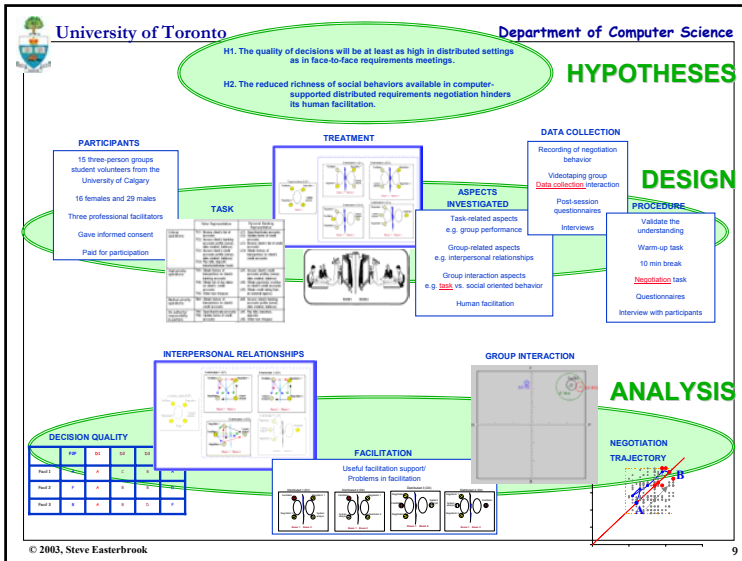
hard to apply if you cannot simulate the right conditions in the lab
limited confidence that the laboratory setup reflects the real situation
ignores contextual factors (e.g. social/organizational/political factors)
extremely time-consuming!

examples

Dana's studies of negotiation settings

See:

D. Perry, A. Porter, L. Votta "Empirical Studies of Software Engineering: A Roadmap". In A. Finkelstein (ed) "The Future of Software Engineering". IEEE CS Press, 2000.



University of Toronto Department of Computer Science

Rational Reconstructions

"a demonstration of a tool or technique on data taken from a real case study, but applied after the fact to demonstrate how the tool/technique would have worked"

good for
 initial validation before expensive pilot studies
 checking the researcher's intuitions about what the tool/technique can do

limitations
 potential bias (you knew the findings before you started)
 easy to ignore "signal-to-noise ratio"

examples
 LAS; BART; ... etc.

See:
 ??

© 2003, Steve Easterbrook 10

University of Toronto Department of Computer Science

Pilot Studies

controlled introduction of a tool/technique into a real project, where the researcher can no longer control the context, but where the net effect can be measured (e.g. against a baseline, or against previous experience)

good for
 can measure the benefits in a real setting
 preparation for tech. transfer
 getting organisations interested in your work

limitations
 hard to get organisations to adopt unproven ideas
 Hawthorne effect (and other bias problems)

examples
 Robyn Lutz has some...

See:
 R. L Glass "Pilot Studies: What, Why and How" J. Systems and Software, vol 36, no 1, pp85-97, 1997

© 2003, Steve Easterbrook 11

University of Toronto Department of Computer Science

Questions

- ⇒ do any of these idioms capture your research?
 - ⊗ do the distinctions make sense?
 - ⊗ are there other idioms we've missed?
- ⇒ Are we (as a community) using the right idioms?
 - ⊗ Should we be using some of them more than we do?
 - ⊗ Should we be using some of them less than we do?
- ⇒ What standards of reporting should we demand?
 - ⊗ e.g. when reviewing papers for the RE conferences
 - ⊗ should we be more explicit about our research methods?
- ⇒ What practical steps can we take...
 - ⊗ workshop at RE'03 on research validation?
 - ⊗ week long workshop in 2004 to define benchmarks?

© 2003, Steve Easterbrook 12



Are we ready for Benchmarks?

- ⇒ **Precondition 1: a minimum level of maturity in the discipline**
 - ↳ increasing concern with validation of research results and comparisons
 - ↳ attempted replication of results
 - ↳ use of proto-benchmarks (or attempts to apply solutions to a common corpus of problems)
 - ↳ increasing resistance to accept speculative papers for publication (willingness to incur the cost of developing and maintaining benchmarks)
 - ↳ (willingness to commit to one particular paradigm)
- ⇒ **Precondition 2: an ethos of collaboration within the community**
 - ↳ willingness to work together on common problems
 - ↳ history of visits between research labs
 - ↳ history of researchers using one another's tools



Benchmark Development

- ⇒ **Consensus building process:**
 - ↳ Effort must be led by a small number of champions
 - ↳ Must be opportunities for the general community to participate and provide feedback
 - ↳ Design decisions for the benchmark need to be supported by laboratory work
- ⇒ **Desiderata for successful benchmarks:**
 - ↳ accessibility
 - ↳ affordability
 - ↳ clarity
 - ↳ relevance
 - ↳ solvability
 - ↳ portability
 - ↳ scalability