1 Bio of presenter

Roel Wieringa (http://www.cs.utwente.nl/ roelw) is Chair of Information Systems and head of the Computer Science Department at the University of Twente, the Netherlands. His research interests include requirements engineering, risk assessment, and research methodology. He has written two books, on Requirements Engineering and on the Design of Reactive Systems. He has been Associate Editor in Chief of IEEE Software for the area of requirements engineering from 2004 to 2007 and serves on the board and program committees of various journals and conferences.

2 Title

Empirical Validation Research Methods

3 Abstract

Empirical validation research is the empirical investigation of properties of newly designed artifacts before they are transferred to practice. An *artifact* is anything designed for a useful purpose, such as a new notation, technique, method, algorithm, device, organization structure or, for that matter, a new medicine. Artifacts are designed and redesigned by researchers, engineers and professionals, and some time after (re)design they may be transferred to practice. An artifact is *transferred to practice* if other people than the designers use it for their own purposes. For example, a software engineering notation has been transferred to practice when companies start using it in their commercial projects; a software algorithm has been transferred to practice when commercially available software contains the algorithm; and a medicine has been transferred to practice when patients buy the medicine and use it to get better.

Empirical *validation* research must be contrasted with the investigation of artifacts when they are used in practice. In this tutorial we refer to this second kind of research as empirical *evaluation* research. For example, we can investigate the way requirements are specified and prioritized in agile projects, or how enterprize architectures are aligned to business goals in practice. In this kind of empirical evaluation research, the object of study is available and can be studied by means of surveys, case studies, and other kinds of field research.

In empirical validation research the artifact is not used in the field yet, and validation takes place by modelling and simulation. For example, in computer science a new lookup algorithm for distributed hash tables in a peer-to-peer network can be validated by (1) building a prototype of the algorithm, (2) building a simulation of the real-world environment in which the algorithm is supposed to be used, and (3) running the prototype against artificial but realistic scenarios in this environment. Similarly, a new technique for prioritizing requirements can be validated by (1) instructing students in the use of the technique, (2) setting up a project in which the students must use the technique,

and (3) providing the student in this project with an artificial but realistic requirements engineering scenario. In both cases, the researcher must assess to which extent the simulation is similar to the intended real-world situation, and to which extent this similarity justifies generalizing from the simulation to the real-world situation.

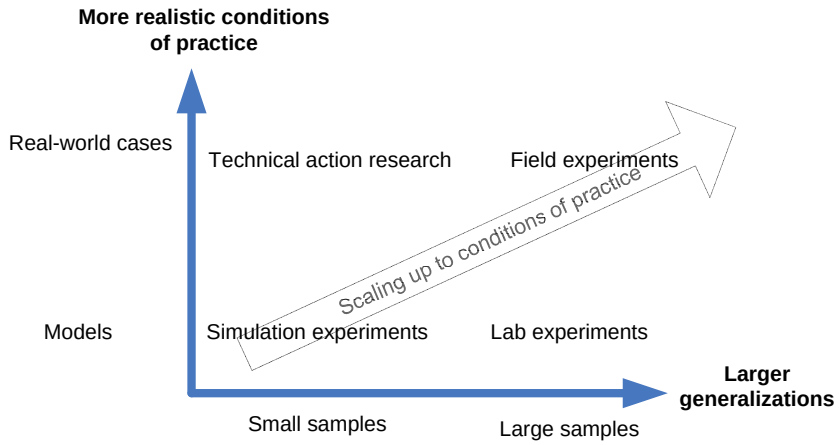
Empirical validation research always proceeds by *scaling up* from artificial laboratory conditions to practical real-world conditions. A newly designed artifact is first tested in idealized and controlled laboratory conditions. After successful laboratory testing, the tests are scaled up in two ways:

1. Addition of realistic conditions of practice.
2. Scaling up to larger sets of subjects.

In information systems and software engineering research, the two dimensions are present too. They can be scaled up separately or simultaneously (see the figure):

- *Simulation experiments.* We can construct a model of the artifact and use it in an idealized, simulated context to study the effects. To scale up the simulation of the real-world environment to conditions of practice, we can remove simplifying assumptions step by step. We can draw conclusions like "in situations relevantly similar to that of the modelled artifact, the artifact will have similar effects."
- *Technical action research.* We can go to the real world rather than working with a model and simulation of it, and use the new artifact ourselves in a real-world project. We can add even more realism by teaching our technique to others and then observing how they use it in practice. We can draw conclusions like "in situations relevantly similar to that of this case, the artifact will have similar effects."
- *Statistical laboratory experiment.* On the other hand, staying in artificial conditions, we can scale up to larger samples by performing a randomized controlled trial on sample of subjects in the laboratory. For example, we can do an experiment with students. We can then draw statistical conclusions of the kind "there is a difference in the laboratory between the treatment group (that used the artifact) and the control group (that used something else)".
- *Statistical field experiment.* Finally, combining real-world conditions with large sample size, we can do field experiments with a sample of software professionals in practice, to draw conclusions like "under realistic conditions, there is a difference between the treatment group (that used the artifact) and the control group (that used something else)."

The ability to scale up along one or both of these dimensions is constrained by the research budget. Field experiments with professionals are extremely expensive and time-consuming and are hardly ever performed. However, simulations,



experiments with students, and action research studies are within the financial and temporal resources of most researchers.

The above research methods are to be contrasted with *illustrations,* in which the researcher uses a small example to illustrate a new artifact, and *laboratory demonstrations,* in which the researcher uses a large, possibly realistic example to check if the artifact *could* be used without bothering to simulate the context of use. Lab demonstrations are often "dry runs", in which the researcher runs through a possible use of the artifact behind his or her desk. The purpose of illustrations and lab demos is to convince the researcher and others that it would be useful to perform empirical validation research, but it does not have the careful design as scientific research has, and therefore cannot support more general conclusions.

After performing an empirical validation, the researcher must draw conclusions about the artifact. The two dimensions of scaling up correspond with two ways of generalizing from an investigation.

1. Scaling up to bigger samples supports statistical inferences;
2. Scaling up to conditions of practice supports case-based inference.

In statistical inference, we draw conclusions about the *difference* that a treatment has compared to another treatment In case-based reasoning, we consider the structure of the case and try to argue that in cases with similar structure, a treatment will have a similar effect.

Whatever reasoning is followed, the conclusions will be fallible and the researcher must indicate the extent of his or her uncertainty about the conclusions. Different research designs entail different threats to validity [3] but one threat shared by all designs is that of lack of similarity: The model, case or subject investigated is not representative of the population to which the treatment will be applied [2]. We will spend some time on the role of representativeness in statistical inference and in case-based inference.

4 Scope

The problem of validating new technology exists in all engineering sciences. The examples of this tutorial are drawn from the fields of information systems and software engineering and, if the audience is interested, from computer science. Examples from information systems include validating new methods and techniques for aligning business and information technology; examples from software engineering include new methods and techniques to develop or maintain software– and information systems. Examples from computer science are typically new algorithms that need to be tested under conditions that simulate practice.

All these fields are currently expanding rapidly in their empirical work. The aim of this tutorial is to contribute to the justifiability and generalizability of knowledge claims that we make in these fields about new technology.

5 Background of the attendees

The intended audience consists of researchers and engineers who want to validate new technology empirically. This includes PhD students and industrial researchers who have designed a new technique, method, software architecture, algorithm or other kind of artifact, and want to show how this new artifact would perform as desired in practice. The audience is assumed to have at least a bachelor-level background in information systems, software engineering or computer science.

6 Material

The attendees will receive a hardcopy of the slides and hardcopies of relevant papers [4, 5, 6].

7 Timetable (90 minutes)

The companion tutorial proposal "Technical Action Research" can be followed independently of this one. Jointly they add up to a 3-hour tutorial starting with this one and continuing with the Technical Action Research tutorial.

- Introduction: Validating new technology (15 minutes)
  - The engineering cycle
  - Validation research questions
  - The fundamental problem of validation
- Scaling up to practice (15 minutes)
  - Scaling up in drug validation research

  - Approaching conditions of practice
  - Scaling up the sample
- Validation research methods (40 minutes)
  - Illustration and laboratory demonstrations
  - Overview of validation research methods
  - Simulation studies
  - Action research
- How to draw conclusions from validation research (15 minutes)
  - Statistical inference
  - Architectural inference
  - Representativeness
- Take-home message (5 minutes)

References

[1] P.B. Seddon and R. Scheepers. Towards the improved treatment of generalization from knowledge claims in IS research: drawing general conclusions from samples. *European Journal of Information Systems*, pages 1–16, 2011. doi:10.1057/ejis.2011.9.

[2] W.R. Shadish, T.D. Cook, and D.T. Campbell. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, 2002.

[3] R. Wieringa, N. Maiden, N. Mead, and C. Rolland. Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Requirements Engineering*, 11(1):102–107, March 2006.

[4] R. J. Wieringa. Design science as nested problem solving. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, Philadelphia*, pages 1–12, New York, 2009. ACM.

[5] R. J. Wieringa. Relevance and problem choice in design science. In *Global Perspectives on Design Science Research (DESRIST). 5th International Conference, St. Gallen*, volume 6105 of *Lecture Notes in Computer Science*, pages 61–76, London, 2010. Springer Verlag.