The Future of Empirical Methods in Software Engineering Research

Dag I. K. Sjøberg, Tore Dybå and Magne Jørgensen

Dag I.K. Sjøberg received the MSc degree in computer science from the University of Oslo in 1987 and the PhD degree in computing science from the University of Glasgow in 1993. He has five years of industry experience as a consultant and group leader. He is now research director of the Department of Software Engineering, Simula Research Laboratory, and a professor of software engineering in the Department of Informatics, University of Oslo. Among his research interests are research methods in empirical software engineering, software processes, software process improvement, software effort estimation, and object-oriented analysis and design. He is a member of the International Software Engineering Research Network, the IEEE, and the editorial board of Empirical Software Engineering.

Tore Dybå received the MSc degree in electrical engineering and computer science from the Norwegian Institute of Technology in 1986 and the PhD degree in computer and information science from the Norwegian University of Science and Technology in 2001. He is the chief scientist at SINTEF ICT and a visiting scientist at the Simula Research Laboratory. Dr. Dybå worked as a consultant for eight years in Norway and Saudi Arabia before he joined SINTEF in 1994. His research interests include empirical and evidence-based software engineering, software process improvement, and organizational learning. He is on the editorial board of Empirical Software Engineering and he is a member of the IEEE and the IEEE Computer Society.

Magne Jørgensen received the Diplom Ingeneur degree in Wirtschaftswissenschaften from the University of Karlsruhe, Germany, in 1988 and the Dr. Scient. degree in informatics from the University of Oslo, Norway in 1994. He has about 10 years industry experience as software developer, project leader and manager. He is now professor in software engineering at University of Oslo and member of the software engineering research group of Simula Research Laboratory in Oslo, Norway. His research focus is on software cost estimation.

Future of Software Engineering(FOSE'07)
0-7695-2829-5/07 $20.00 © 2007 IEEE

The Future of Empirical Methods in Software Engineering Research

Dag I. K. Sjøberg
Simula Research Laboratory
P.O. Box 134, NO-1325
Lysaker, Norway
dagsj@simula.no

Tore Dybå
Simula Research Laboratory
P.O. Box 134, NO-1325
Lysaker, Norway and
SINTEF ICT, NO-7465
Trondheim, Norway
tore.dyba@sintef.no

Magne Jørgensen
Simula Research Laboratory
P.O. Box 134, NO-1325
Lysaker, Norway
magnej@simula.no

Abstract

We present the vision that for all fields of software engineering (SE), empirical research methods should enable the development of scientific knowledge about how useful different SE technologies are for different kinds of actors, performing different kinds of activities, on different kinds of systems. It is part of the vision that such scientific knowledge will guide the development of new SE technology and is a major input to important SE decisions in industry. Major challenges to the pursuit of this vision are: more SE research should be based on the use of empirical methods; the quality, including relevance, of the studies using such methods should be increased; there should be more and better synthesis of empirical evidence; and more theories should be built and tested. Means to meet these challenges include (1) increased competence regarding how to apply and combine alternative empirical methods, (2) tighter links between academia and industry, (3) the development of common research agendas with a focus on empirical methods, and (4) more resources for empirical research.

1. Introduction

Software systems form the foundation of the modern information society, and many of those systems are among the most complex things ever created. Software engineering (SE) is about developing, maintaining and managing high-quality software systems in a cost-effective and predictable way. SE research studies the real-world phenomena of SE and concerns (1) the development of new, or modification of existing, technologies (process models, methods, techniques, tools or languages) to support SE activities, and (2) the evaluation and comparison of the effect of using such technology in the often very complex interaction of individuals, teams, projects and organisations, and various types of task and software system. Sciences that study real-world phenomena, i.e., empirical sciences, of necessity use empirical methods, which use consists of gathering information on the basis of systematic observation and experiment, rather than deductive logic or mathematics. Hence, if SE research is to be scientific, it too must use empirical methods.

Activities (1) and (2) are mutually dependent, and both are crucial to the success of SE practice. Historically, activity (1) seems to have been emphasised.

An empirical approach to assessing SE technology, including industrial collaboration, started on a large scale in the 1970s with the work of Vic Basili and his group at the University of Maryland [9, 19]. Since then, there has been an increased focus on the need for, and approaches to, applying empirical methods in SE research [10, 11, 101, 113, 129, 130, 136]. The focus on empirical SE is reflected in forums such as the Journal of Empirical SE (EMSE, from 1996), IEEE International Symposium on Software Metrics (METRICS, from 1993), Empirical Assessment & Evaluation in SE (EASE, from 1997) and IEEE International Symposium on Empirical SE (ISESE, from 2002).

Despite the increased focus, we remain far from our vision (Section 3). Attaining the vision will require more empirical studies (of higher quality, including relevance, than at present), and more focus on synthesizing evidence and building theories (Section 4). There are several ways of addressing these challenges: increasing competence regarding how to conduct empirical studies; improving the links between academia and industry; developing common research agendas, which would also increase the credit given for conducting high-quality, often very time-consuming, empirical studies; and increasing the resources available for such work to an extent commensurate with the importance of software systems in society (Section 5).

Future of Software Engineering(FOSE'07)
0-7695-2829-5/07 $20.00 © 2007 IEEE

2. Empirical methods

This section describes the scientific method and provides an overview of the empirical research methods and terminology most relevant to SE, as well as some suggestions for further reading.

2.1 The scientific method

Empirical science concerns the acquisition of knowledge by empirical methods. However, what constitutes knowledge, and hence the methods for acquiring it, rests on basic assumptions regarding ontology (i.e., what we believe to exist) and epistemology (i.e., how beliefs are acquired and what justifies them). Further, scientific method is not monolithic, but is constituted by the concepts, rules, techniques, and approaches that are used by a great variety of scientific disciplines. The inductive and hypothetico-deductive methods are commonly regarded as the two main theories of scientific method, but there are also broader theories that are based on abduction; see, e.g., [55].

Empirical research seeks to explore, describe, predict, and explain natural, social, or cognitive phenomena by using evidence based on observation or experience. It involves obtaining and interpreting evidence by, e.g., experimentation, systematic observation, interviews or surveys, or by the careful examination of documents or artifacts.

Approaches to empirical research can incorporate both qualitative and quantitative methods for collecting and analyzing data. *Quantitative methods* collect numerical data and analyze it using statistical methods, while *qualitative methods* collect material in the form of text, images or sounds drawn from observations, interviews and documentary evidence, and analyze it using methods that do not rely on precise measurement to yield their conclusions; see, e.g., [57, 64, 87, 93, 98, 111, 116].

Although different approaches to research suggest different steps in the process of acquiring knowledge, most empirical methods require that the researcher specify a research question, design the study, gather the data or evidence, analyze the data, and interpret the data. Others are then informed about the newly acquired knowledge.

There are also a number of general *design elements* that can be used to strengthen an empirical study by reducing the number and plausibility of internal threats to validity. Shadish *et al*. [118] placed these elements into the following groups: (1) assignment, (2) measurement, (3) comparison groups, and (4) treatments.

2.2 Primary research

We now provide an overview of the most common primary approaches to research in SE. Such *primary research* involves the collection and analysis of original data, utilizing methods such as experimentation, surveys, case studies, and action research.

2.2.1 Experimentation. An experiment is an empirical inquiry that investigates causal relations and processes. The identification of causal relations provides an explanation of *why* a phenomenon occurred, while the identification of casual processes yields an account of *how* a phenomenon occurred [134]. Experiments are conducted when the investigator wants control over the situation, with direct, precise, and systematic manipulation of the behavior of the phenomenon to be studied [133].

All experiments involve at least a treatment, an outcome measure, units of assignment, and some comparison from which change can be inferred and (hopefully) attributed to the treatment. *Randomized (or true) experiments* are characterized by the use of initial random assignments of subjects to experimental groups to infer treatment-cause change. *Quasi-experiments* also have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred. Instead, the comparisons depend on non-equivalent groups that differ from each other in many ways other than the presence of a treatment whose effects are being tested. The task of interpreting the results from a quasi-experiment is, thus, basically one of separating the effects of a treatment from those due to the initial incomparability between the average units in each treatment group, since only the effects of the treatment are of research interest [30].

While experiments can help to provide inductive support for hypotheses, their most important application is in *testing* theories and hypotheses. If an experiment uncovers a single instance of an event that contradicts that which is predicted by an hypothesis or theory, the hypothesis or theory may be rejected [107]. Note that in social and behavioral sciences, with which empirical SE shares many methodological issues, deeming a theory as false based on its predictions is rarely feasible [88, 132]. If a prediction is not supported by empirical evidence, alternative theories or refinements of existing theories are sought, rather than theory rejection. SE experiments are typically used to explore relationships among data points describing one variable or across multiple variables, to evaluate the accuracy of models, or to validate measures [123].

Experiments can be differentiated according to the

degree of realism in the research setting [8, 120, 121]; i.e., laboratory experiments and field experiments.

General guidelines for experimental design and analysis can be found in [26, 30, 95, 118]. Specific guidelines for conducting SE experiments can be found in [11, 71, 104, 133]. Furthermore, Kitchenham *et al*. [77] have proposed preliminary guidelines for empirical research in SE that are well-suited for experimental research. Overviews of published SE experiments can be found in [11, 123, 139].

2.2.2 Surveys. A survey is a retrospective study of a situation that investigates relationships and outcomes. It is useful for studying a large number of variables using a large sample size and rigorous statistical analysis. Surveys are especially well-suited for answering questions about what, how much, and how many, as well as questions about how and why [106]. They are used when control of the independent and dependent variables is not possible or not desirable, when the phenomena of interest must be studied in their natural setting, and when the phenomena of interest occur in current time or the recent past.

Conducting surveys is a standard method of empirical study in disciplines such as marketing, medicine, psychology, and sociology. There is also a long tradition for the use of surveys as an intervention strategy for organizational change; see, e.g., [81, 97]. In SE, surveys usually poll a set of data from an event that has occurred to determine how the population reacted to a particular method, tool, or technique, or to determine trends or relationships. They try to capture what is happening broadly over large groups of projects.

Survey designs may be categorized as cross-sectional or longitudinal, depending on whether they exclude or include explicit attention to the time dimension [106].

A general review of survey research from the vantage point of psychology is provided by [82]. General introductions and guidelines for survey research can be found in [47, 98], while an assessment of survey research in management information systems can be found in [106]. Details regarding instrument design and scale development are given in [35, 124]. An example of the construction of an instrument for SE survey research can be found in [39].

2.2.3 Case studies. A case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident [134]. So, while an experiment deliberately divorces a phenomenon from its context and a survey's ability to investigate the context is limited, the case study aims deliberately at covering the contextual conditions.

Yin [134] noted that a case study has a distinct advantage when a 'how' or 'why' question is being asked about a contemporary set of events over which the investigator has little or no control. For SE, case studies are also useful in answering a "which is better" question [78].

In general, case study designs can be *single-case* or *multiple-case* studies, and they can involve a single unit (holistic) or multiple units (embedded) of analysis [134]. There are, thus, four general designs for case studies: (1) single-case, holistic design, (2) single-case, embedded design, (3) multiple-case, holistic design, and (4) multiple-case, embedded design.

Conducting case studies is a standard method of empirical study in management and related disciplines such as organization development and information systems (IS) research.

In SE, case studies are particularly important for the industrial evaluation of SE methods and tools, because they can avoid the scale-up problems that are often associated with experiments. To avoid bias and ensure internal validity, it is necessary to identify a valid basis for assessing the results of the case study. Basically, there are three ways of designing an SE case study to facilitate this [78]: results can be compared with a company baseline, with a sister project, or components within a project can be compared.

Standard texts on case study research include [48, 126, 134]. A discussion of the case study research strategy in studies of information systems can be found in [17, 85]. Guidelines and checklists for case studies in SE can be found in [50, 74, 78].

2.2.4 Action research. Action research focuses particularly on combining theory and practice [54]. It attempts to provide practical value to the client organization while simultaneously contributing to the acquisition of new theoretical knowledge. It can be characterized as "an iterative process involving researchers and practitioners acting together on a particular cycle of activities, including problem diagnosis, action intervention, and reflective learning." [6, p. 94]. The major strength of action research is, thus, the in-depth and first-hand understanding the researcher obtains. Its weakness is the potential lack of objectivity on the part of the researchers when they attempt to secure a successful outcome for the client organization [17].

Three distinct types of model can be seen in action research: iterative, reflective, and linear [15].

General introductions to action research and its cognates can be found in [44, 110]. A general discussion of the applicability of action research to IS research is

provided by [6, 80], specific frameworks for action research in IS are presented by [15, 34, 84], while a critical perspective on action research as a method for IS research can be found in [14]. Principles for conducting and evaluating interpretative research in IS are proposed in [79].

2.3 Secondary research

Secondary research uses data from previously published studies for the purpose of *research synthesis*, which is the collective term for a family of methods for summarizing, integrating and, where possible, combining the findings of different studies on a topic or research question [31]. Such synthesis can also identify crucial areas and questions that have not been addressed adequately with past empirical research. It is built upon the observation that no matter how well-designed and executed, empirical findings from single studies are limited in the extent to which they may be generalized [27].

Research synthesis is often used interchangeably with "systematic review" and "systematic literature review". The strength of these methods lies in their explicit attempts to minimize the chances of drawing incorrect or misleading conclusions as a result of biases in primary studies or from biases arising from the review process itself. They are essential for informing research and practice, and most of the current interest in systematic reviews within SE and other disciplines originates from reviews of the effectiveness of interventions reflected in initiatives such as the Campbell (www.campbellcollaboration.org) and Cochrane (www.cochrane.org) Collaborations.

Systematic reviews are one of the key building blocks of evidence-based SE [41, 70, 76], and the interest in conducting such reviews within SE is clearly growing [40, 56, 65, 66, 123], although the coverage of SE topics by systematic reviews is still in its infancy and very limited.

If primary studies are similar enough with respect to interventions and outcome variables, it may be possible to synthesize them by meta-analysis, which uses statistical methods to combine effect sizes. However, in SE, primary studies are often too heterogeneous to permit a statistical summary and, in particular, for qualitative and mixed methods studies, different methods of research synthesis are required, e.g., *meta-ethnography* [99] or *meta-study* [100].

Standard texts on research synthesis can be found in [31, 32, 43, 96, 102]. General introductions to meta-analysis are provided in [90, 114], while the question of how systematic reviews can incorporate qualitative research is addressed in [36, 37, 127]. General procedures for performing systematic reviews in SE have been proposed in [76]. Guidelines for the use of meta-analysis in SE can be found in [94, 105].

3. Vision

Regarding the role of empirical methods in SE in the future (2020-2025), our vision is as follows:

> *In all fields of SE, empirical methods should enable the development of scientific knowledge about how useful different SE technologies are, for different kinds of actors, performing different kinds of activities, on different kinds of systems. Such scientific knowledge should guide the development of new SE technology and be a major input to important SE decisions in industry and services.*

A particular challenge when describing scientific knowledge is to identify the appropriate level of abstraction. The level will depend on how the knowledge is used. Given such a level, corresponding taxonomies should also be developed. The typical SE situation is that an *actor* applies *technologies* to perform certain activities on an (existing or planned) *software system*. These high-level concepts or "archetype classes" [122] with typical sub-concepts or subclasses are listed in Table 1. One may also envisage collections of (component) classes for each of the (sub)classes. For example, component classes of a software system may be requirement specifications, design models, source and executable code, test documents, various kinds of documentation, etc.

In addition, appropriate characteristics of the classes, and their relative effect, should also be identified and measured. For example, the usefulness of a technology for a given activity may depend on characteristics of the software engineers, such as their experi-

Table 1. Archetype classes

Archetype Class	Subclasses
• Actor	• individual, team, project, organisation or industry
• Technology	• process model, method, technique, tool [138] or language
• Activity	• plan, create, modify or analyze (a software system); see also [123]
• Software system	• software systems may be classified along many dimensions, such as size, complexity, application domain [52], business/scientific/student project or administrative/embedded/real time, etc.

ence, education, mental ability, personality, motivation, and knowledge of a software system, including its application domain and technological environment. Note that contexts or environments are supposed to be part of the descriptions of the respective archetype classes.

An important aspect of our vision is the *size of an effect*, which is vital for judging practical importance. In many cases, it will be trivial that a certain technology has a positive effect in a given context, but if the effect can be *quantified* (somewhat), the importance of the knowledge for decision-making will, in most cases, increase significantly, although it may be far from trivial to quantify the effect at a general level.

Our vision includes the hope that scientific knowledge will guide the development of new, and modification of existing, SE technology. Theories that predict the effect of new or modified technology before it is in widespread use by the software industry would be particularly useful. At present, the adoption of a technology may be attributed to many factors besides effectiveness. Achievements in this area will demand more collaboration between those who do development research and those who do evaluative research.

We believe that the SE community will observe higher efficiency, better quality and more rational investments as a consequence of making decisions based on scientific knowledge rather than on personal opinion. Stakeholders here are not only people within the SE community, but also include, for example, governments, clients and software users. Note that our vision does not imply that expert judgment in SE decisions becomes unimportant. On the contrary, good expert judgment will be essential for guiding the identification of high-quality studies, including the evaluation of their usefulness, and the transference of results to the context in which they will be used. In addition, expert judgment may be necessary to guide when to investigate something empirically and when qualified opinions are likely to be of sufficient quality. This type of decision will be affected by, among other factors, time pressure, importance, available resources, and cost-benefit considerations.

Sections 4 and 5 describe, respectively, some challenges to the pursuit of our vision and some ways of addressing them.

4. Challenges

The suggested changes related to empirical methods, the motivation for the changes, the means to complete the changes and possible threats are included in the discussion of each of the challenges.

4.1 More empirical SE studies

How many empirical SE studies do we need? There is, of course, no exact answer to that question, i.e., it is an example of a poorly formulated research question (see Section 4.3). Nevertheless, for purposes of illustration, let us try to calculate the number of studies required:

> *Assume that there are 1000 research questions of high industrial importance (the real number obviously depends on how we define "research question" and "high industrial importance") that it is meaningful to decide empirically. Furthermore, assume that a review of empirical studies related to a research question requires at least 20 high quality studies, conducted over the last 10 years, to derive a medium to strong support for particular SE practices in the most relevant development contexts. Even this scenario, which is probably highly optimistic, requires that we conduct at least 2000 high-quality empirical studies every year.*

What is the current situation with respect to the number of SE studies? From of a total of 5453 scientific articles published in 12 major SE journals and conferences in the decade 1993–2002, Sjøberg *et al*. [123] identified 113 *controlled experiments* in which humans performed SE tasks. They were reported in 103 (1.9%) articles. Glass *et al*. [51] and Zelkowitz and Wallace [137] reported about 3% controlled experiments. There

Table 2. Extent of empirical studies

State of Practice	Target (2020-2025)
• There are relatively few empirical studies. The focus on evaluation of technology is lower than that of developing new technology	• A large number of studies covering all important fields of SE and using different empirical methods are conducted and reported. Most research that leads to new or modified technology is accompanied with empirical evaluation. At least for journal papers, there should be good reasons for not including a proper evaluation.
• The use of empirical methods by the software industry is low.	• Most large software development organizations have personnel highly skilled in designing and conducting empirical studies to support their own decisions.

were very few experiments in the period 1993-1995. The number rose in 1996, but there was no increasing trend from 1996 to 2002 [123].

Regarding *surveys* as the primary research method, we have found only one relevant literature review; Glass *et al*. [51] classified 1.6% of their 369 papers as "descriptive/explorative survey".

Case studies were identified by Glass *et al*. [51] in 2.2% (8 of 369) of the papers; Zelkowitz and Wallace [137] found 10.3% (58 of 612 papers). Similarly to the review of controlled experiments, Simula Research Laboratory has begun work on a review of case studies in SE. In a pilot study [63], we identified 12% case studies (50 of 427 articles that were randomly selected from the 5453 articles mentioned above). One reason for the large differences in the proportion of case studies may be due to the difficulty in defining and identifying case studies. In our review of controlled experiments, we had an operational definition, but in our case study pilot we simply based the identification on what the authors themselves said. This means that, for example, demonstrations of the use of technology, typically performed by the authors, were included if the authors called it a case study. This applied to 58% of the papers. If we included only "evaluative case studies", we would end up with only 4.9% case studies.

Glass *et al*. [51] found no *action research* studies. Based on the search term "action research", we found 14 articles out of 6749 in articles published in the nine journals and three conferences described in [123] in the period 1993-2006. Ten of the articles were published in 2003 or later, so there might be an increase here. Note that, as for case studies, the understanding of what action research means in the context of SE is little understood. So, there may be studies that are reported as case studies, while they in fact are action research studies as defined in Section 2.2.4, and vice versa; some of the 14 studies reported as action research may in fact not meet the criteria defined in Section 2.2.4. In any case, action research seems almost absent in the SE literature.

Regarding *secondary research*, literature reviews and meta-analyses were identified in 1.1% of the papers by Glass *et al* [51] and 3.0 % by Zelkowitz and Wallace [137].

Regarding empirical studies as a whole, Tichy *et al*. [130] reported a proportion of 17% (15 of 87 IEEE TSE papers). Glass *et al*. [51] characterized 14% (51 of 369) as "evaluative". Our own review work (2% experiments, 5% or 10% case studies (depending on definition), 0% action research indicates the same level. Rather different are the findings by Zelkowitz and Wallace [137]. They classify two-thirds as having "experimental evaluation", but about half of these are "assertive", which is similar to "demonstrative" case studies as described above.

The reason for differences found in the review references above may be that there are differences in the number of sources, differences in the definition, and hence inclusion criteria, of the various study types, and differences in the publication year of the different articles. Nevertheless, an average of the reviews indicates that about 20% of all papers report empirical studies. Assume further that there are 1.1 studies for each paper (the number found in [123]. Then there would be 5453 * 20% * 1.1 = 1200 studies published in the 12 SE journals and conferences in the decade 1993–2002 described in [123]. Assume further that there are another 50% (a high estimate) of studies that are reported in other SE journals and conferences. Then the total would be 1800 studies, that is, 180 a year. Of course, there are many uncertainties in this estimate, but it is still another order of magnitude lower than the speculated level needed to reach our vision of 2000 a year. Note also that those 2000 studies were supposed to have high quality, including being relevant. In the estimate of actual production of 180 a year, all published studies are included. Consequently, our vision depends on a substantial increase in the number of empirical studies conducted; see Table 2.

4.2 Increased quality of empirical studies

To achieve the vision of the development of scientific knowledge and empirically-based SE decisions, we need studies that we can trust, i.e., studies with high *validity*. There are numerous textbooks on how to design and conduct high quality research; see Section 2 for a brief overview. This section focuses on a few, non-trivial, important empirical method challenges in SE; see Table 3.

The goal of empirical SE studies is not to maximize their quality, but to find the right level of quality. The right level of quality depends on, among other issues, the importance of not drawing incorrect conclusions and the cost involved in increasing quality. Lack of knowledge about the right level of quality may easily be used as an excuse to conduct low quality studies. It might be said, for example, that even a low quality study is better than no study at all. This is a dangerous attitude that may have as consequences that: i) incorrect results are spread, ii) people's trust in the results of empirical studies will be eroded, and, iii) it will be difficult to create a culture of quality among the people using empirical software methods.

Determining the right level of quality is difficult. Different researchers may assess the level of quality of the same study differently, the importance of a study is

frequently not known when design decisions are taken, the amount of resources required limits the levels of quality achievable, and there is an inherent conflict between the internal and external validity of studies. In addition, if we require a level of study quality that is too high, fresh researchers may be discouraged from pursuing a research career due to a high frequency of rejected papers. Probably, we will have to live with a substantial proportion of low quality and/or irrelevant research studies. This is the price we must pay to educate new researchers and allow studies with innovative research designs, which have a high risk of finding nothing useful.

A particularly important quality threat within empirical SE studies is related to construct validity; see e.g., [42]. On the one hand, we need to measure something in order to understand it, e.g., we need to measure code maintainability in order to understand how it relates to code complexity. On the other hand, and just as importantly, we need to understand something in order to measure it. We have the impression that many studies try to measure phenomena that are poorly understood, with the consequence that the construct validity is low. As an illustration, there are SE studies that attempt to measure "software quality" and apply a measure based on faults per lines of code. Typically, the relation between the construct (quality) and the measure (faults per lines of code) is not established; it is based solely on the vague idea that fault density has something to do with quality [69]. There is, consequently, both a construct validity problem and a problem of precision (and perhaps honesty) present, i.e., why call something "software quality" when the measured phenomenon may more precisely be termed "fault density"? Higher quality studies are only possible if we use constructs that we understand well and are able to communicate precisely.

To produce knowledge that applies to a wide range of settings, we must compare and integrate the results from different studies and generalize them to settings beyond the original studies. To produce general knowledge, we must know the *scope of validity* of the results from the studies. In SE, the scope of validity is the populations of, for example, actors, technologies, activities, and software systems (see Section 3) for which the results are valid. In the study of 113 experiments [123], the target populations studied were stated in only two cases. Most papers reported a very wide, implicit notion of scope in a section on "threats to external validity". For example, it is typically discussed whether the results from an experiment that used students as subjects are applicable to "professionals" without characterizing the kind of professional in mind, e.g., skills, education and experience in general and specifically to the technology investigated in the experiment. Possible consequences of this lack of awareness regarding the definition and reporting of the scope of validity or of defining very wide scopes are as follows:

- Many apparently conflicting results are reported.
- Interpreting the applicability of the results of an experiment becomes difficult because of many confounding factors (narrower scopes will generally result in fewer confounding factors).
- Random sampling, replication and research synthesis become difficult.

Hence, the scope of studies should be defined and reported systematically and explicitly, and it is a good idea to formulate the scope relatively narrowly to begin with and then extend it gradually through various kinds of replications. Building knowledge and theories in a careful bottom-up fashion, in which the scopes are gradually extended, is done successfully in other sciences. For example, the *prospect theory* developed by Kahneman and Tversky [72] had initially a very narrow scope; the theory was only applied to simple gambling problems. It took them, and other researchers, another thirteen years to extend the theory to make it more generally applicable. In 2002, they received the Nobel Prize in economics for this work. The scope of a hypothesis or theory is expanded when studies are replicated on new populations and in new contexts. It may require significant effort by several research groups over many years to provide validated results that cover (a substantial part of) the scope of interest on a given topic.

Note that, to avoid bias, replications should preferably be conducted by others than those conducting the original experiment. Among the 15 *differentiated replications* [89] identified in [123], seven of eight replications carried out by the same authors confirmed the results of the original experiments, while only one of seven replications carried out by other researchers confirmed the original results. Studies that evaluate the ability to learn from experience have demonstrated biases that prevent people from using the information provided by such experience. Such biases include preferences for confirmatory evidence, assumptions about causality, and a disregard of negative information [21]. When evaluating their predictions, people have difficulty in searching for information that may count against them. This issue pertains to expectation-based experimentation as a whole, but might be particularly relevant when replicating one's own studies.

The empirical SE community has focused on how to conduct controlled experiments and the number of reviews has increased over the last few years. Even

though there is still a need to improve the way in which experiments are conducted, it may be even more important to focus on how to achieve better surveys, case studies and action research. For example, most studies termed "case studies" lack systematic observations and analyses of the use of the technology, and they should preferably be carried out in an industrial setting. We share the opinion of Yin, who states [134, p. 17]:

> *… we must all work hard to overcome the problem of doing case study research, including the recognition that some of us were not meant by skill or disposition, to do such research in the first place. Case study research is remarkably hard even though case studies have traditionally been considered 'soft' research, possibly because investigators have not followed systematic procedures.*

Having said this, it is our experience that some reviewers have very little knowledge of types of empirical study other than experiments. For example, they are only aware of *statistical* generalisation and thus criticize case studies as having only one sampling unit. Yin has met this problem in the disciplines from which he has experience [134, p. 32]:

> *A fatal flaw in doing case studies is to conceive of statistical generalization as the method of generalizing the results of the case study. This is because your cases are not "sampling units" and should not be chosen for this reason. Rather, individual case studies are to be selected as a laboratory investigator selects the topic of a new experiment. Multiple cases, in this sense, should be considered like multiple experiments.*

Table 3. Quality of empirical studies

State of Practice	Target (2020-2025)
• Researchers frequently do not build sufficiently on previous research results, particularly those achieved outside the researcher's own domain.	• There is a strong emphasis on building on previous research results, including those from other disciplines.
• Research method and included design elements are frequently applied without careful consideration of alternative study designs. Skills in conducting controlled experiments and reviews seem to have improved over the last few years, but not skills in conducting surveys, case-studies and action research.	• Research method and design elements are carefully selected and combined, based on an in-depth understanding of their strengths and weaknesses. Researchers are trained in using a large set of research methods and techniques. Skills in conducting case studies and action research are stimulated by the work in the IS field.
• Study results are frequently not robust due to lack of replications and use of only one type of research design.	• Replications and triangulation of research designs are frequently used means for achieving robust results.
• Studies are frequently conducted by researchers with a vested interest in study outcome, with insufficient precautions to prevent biases. Many of these studies are merely demonstrations that a technology works ("proof of concept") or simple experience reports ("lessons learned").	• Empirical evaluation is mainly based on high quality studies conducted by researchers with no vested interest in the study outcome.
• Reference points for comparisons of technologies are frequently not stated, or not relevant.	• New technology is compared with relevant alternative technology used in the software industry.
• The scope of validity of empirical studies is rarely defined explicitly.	• The scope is systematically and explicitly defined and reported; it is relatively narrow to begin with and then gradually extended through replications.
• Statistical methods are used mechanically, and with little knowledge about limitations and assumptions. In particular, populations are not well defined. Moreover, for experiments there is a lack of power analysis [40] and effect size estimation [73].	• The use of statistical methods is mature. Populations are well defined, and power analysis and effect size estimation are conducted when appropriate.
• Statistics-based generalization is the dominant means of generalization.	• SE studies include a diverse and reflected view on how to generalize, particularly through the use of theory.

4.3 Increased relevance of empirical studies

Since the ultimate goal of SE research is to support practical software development, we will discuss the quality attribute *relevance* separately. Nevertheless, since there are many types of user and use of research results, relevance should not be interpreted narrowly. Relevance also concerns the indirect usefulness of results in an industrial context, e.g., by providing study results that enable the development of guidelines, and indirect usefulness in a research context, e.g., through a study's impact on other studies or a better understanding of important phenomena through theory building. Note also that our request for greater relevance should not totally exclude studies of an exploratory nature and studies where the likelihood of finding anything useful is very low. The history of science includes many examples of studies where the scope of the study's relevance was first understood by other researchers many years after the research was conducted. As an illustration, essential features of general purpose object-oriented languages today are based on research aiming at the development of a special purpose programming language for simulating discrete event systems: Simula [33].

Returning to industrial relevance, one may divide the relevance of a study into the *investigated topic* and the *implications of the results* [16]. Due to the limited extent of empirical research in SE (Section 4.1), there are obviously many topics that have not been studied. Höfer and Tichy [62] state that the range of software topics studied empirically is rather narrow and that several important topics receive no attention whatsoever. An overview of topics investigated in SE experiments can be found in [123].

In general, the more similar the research setting of a study is to the setting in which the results will be applied, the more directly relevant the study is perceived. Fenton *et al*. [46] state that "evaluative research must involve realistic projects with realistic subjects, and it must be done with sufficient rigor to ensure that any benefits identified are clearly derived from the concept in question. This type of research is time-consuming and expensive and, admittedly, difficult to employ in all software-engineering research. It is not surprising that little of it is being done." See also [58, 119].

The most realistic research setting is found in action research studies, because the setting of the study is the same as the setting in which the results will be applied for a given organization, apart from the presence of the researchers(s). The setting of industry-based case studies is also generally very similar to the setting of application, although researchers *may* study phenomena that might not be regarded as very relevant by the studied organization. Hence, more (high-quality) action research and case studies should be conducted.

Nevertheless, generalizing the results from action research and case studies to settings beyond the studied organizations is a challenge (cf. the discussion of generalization in Section 4.2). The classical method of identifying general cause-effect relationships is to conduct experiments, but how do we make their results relevant to the software industry? Below, we will discuss the realism of experiments relative to the four archetype classes Actor, Technology, Activity and Software system (Table 1).

In SE experiments, the actors are mostly individuals, but sometimes teams. About 90% of the subjects who take part in the experiments are students [123]. Experiments with a large number of subjects have shown significant difference between categories of subjects *with respect to the technology that was most beneficial* [3, 4]. Note that this is an illustration of the fact that many aspects of the complexity of SE only manifest themselves in controlled experiments if the scale is sufficiently large. In the study reported in [4], the analysis was based on three variables: pair programming (two levels), control style (two levels) and programmer category (three levels), resulting in 12 levels (or groups). The power analysis showed that we needed N = 14 observations in each of the 12 groups, for a total N = 168, to get a minimum power of 0.8 for all three main effects and interactions.

The applicability of most experimental results to an industrial setting may, therefore, be questioned. Of course, there may still be good reasons for conducting experiments with students as subjects, such as for testing experimental design and initial hypotheses, or for educational purposes [128]. Note that this issue is more subtle than merely whether one should use professionals or students as subjects. There may be large differences among categories of professionals, and of categories of students. The point is that the skill or expertise level *relative to the technology being evaluated* must be made explicit, thus indicating the population to which the results apply. The typical method by which the expertise level is revealed before an experiment is to run pre-tests [118]. Of course, if one adheres to the principle of random sampling from well-defined populations (see Section 4.2), one does not need to run pre-tests; but this has seemed unrealistic in SE experimentation for many years.

The relevance of the technology being evaluated in an experiment relates to the issue of topic of experiment discussed above. Another dimension is the realism of the technological environment of the experiment. There is relatively little reporting of the impact of technology environment on experimental results,

but it seems clear that artificial class room settings without professional development tools may, in many situations, threaten the validity of the results. Hence, one should attempt to make the technological environment as realistic as possible if it is possible that it will influence the results, even though extra effort is needed to conduct experiments in realistic settings [121].

SE activities are constituted by *tasks*, which typically have time limits. Large development tasks may take months, while many maintenance tasks may take only a couple of hours. Nevertheless, the typical task in SE experiments are much smaller than typical industrial tasks; in the survey reported in [123], the median duration of experiments was 1.0 hour for experiments in which the time of each subject was taken and 2.0 hours when only an overall time for the whole experiment was given.

Most software systems involved in SE experiments are either constructed for the purpose of the experiment or are student projects; in the review reported in [123], only 14% were commercial systems. Accordingly, the systems are generally small and simple. The experiment reported in [4] demonstrated that system complexity has an impact; the system complexity had a significant effect on the effect of pair programming with respect to producing correct programs.

About half of the of articles reported in [123] mention that the tasks and systems used in the experiments are not representative of industrial tasks and systems with respect to size/duration, complexity, application domain, etc. A few articles claim that there is no threat to external validity because the experiment was conducted in an industrial context; they used, for example, industrial software.

The community seems to agree that it is a problem that most experiments do not resemble an industrial situation, but one challenge is to define what an industrial situation is. There are an endless number of industrial technologies, tasks and systems, so what is a representative technology, task or system? First, well-defined taxonomies are needed. Then, the representativeness of the categories of the taxonomies can be determined by, for example, conducting surveys, logging the activities in certain companies, or consulting project information databases. Afterwards, experiments could sample from the populations indicated by the categories of these taxonomies.

Finally, the relationships among the four archetype classes should be investigated. For example, a professional development tool will probably become more useful the larger and more complex the tasks and systems become, assuming that the subjects are sufficiently proficient with the tool.

Table 4 summarizes what has been written above, and includes other important changes that we believe are needed for increasing the relevance of empirical SE studies.

4.4 Synthesizing evidence

Science is cumulative. Therefore, we believe that stakeholders who wish to use research to inform their decisions will insist on research syntheses, rather than on the results of small, standalone studies that are currently available to consumers of SE research. A central challenge is thus to provide appropriate methods for synthesizing evidence from diverse types of study and to establish common research agendas within relevant research areas. This challenge calls for several changes related to identifying and selecting primary studies, assessing the quality and scope of primary studies, and synthesizing the results of heterogeneous studies; see Table 5.

4.4.1 Identifying and selecting primary studies. Finding literature for systematic reviews requires identifying all relevant sources of studies and executing a comprehensive search strategy. SE reviewers should consider using multiple digital libraries that include reports from both SE and related disciplines [41].

Due to the lack of standardization among the various electronic resources, developing a search strategy and the selection of search terms requires careful thought. This is because concepts, subject descriptors, and keywords vary among the digital libraries, which is partly a reflection of the lack of common terminology within the SE community itself. Some topics in SE research do not map well to SE subject descriptors and keywords. Hence, an iterative approach using several related terms is often required when an unfamiliar library is being used or when a new topic is being researched.

Additional challenges to undertaking systematic reviews within the domain of SE include the limited facilities offered by the SE-specific bibliographic databases (e.g., for advanced Boolean searches and for downloading citations with abstracts into bibliographic management programs). A challenge is to enhance the electronic SE resources to include functionality to better support the identification, selection, analysis and retrieval of bibliographic and full-text information about SE research.

Table 4. Relevance of empirical studies

State of Practice	Target (2020-2025)
• One may question the industrial relevance of most SE studies.	• More case studies and action research should be carried out. Experiments should show more realism regarding subjects, technology, tasks and software systems.
• Few results answer questions posed by industrial users, e.g., the question "Which method should we use in our context?" It is of little relevance whether method *X* or method *Y* is better with respect to one property in a context with unknown characteristics. The current focus is frequently on comparing mean values of technologies without a proper understanding of individual differences or the studied population.	• A larger part of the research synthesizes and presents results so that it is possible for industrial users to apply them, e.g., through checklists and guidelines. This may include a stronger focus on individualized results, individual differences, and better descriptions of populations and contexts (see Section 4.2); it may be highly relevant to know why, when and how method *X* is better than method *Y* in a carefully selected context with known characteristics.
• Few studies provide results that enable efficient cumulative research, or that are highly relevant for other researchers. An illustration is when we produce studies that compare methods *X* and *Y* with diverging results without explaining the reasons for the difference.	• More research studies are designed with the goal of enabling efficient use of its results by other researchers.
• Important results are hidden in academic language and mathematical notation, and thus not transferred to potential users.	• More focus on communicating important results in plain language in channels where the software industry collects information.

Table 5. Synthesis of evidence

State of Practice	Target (2020-2025)
• Narrative, biased reviews and little appreciation of the value of systematic reviews.	• Scientific methods are used to undertake integrative and interpretive reviews to inform research and practice. Systematic reviews are solicited by all scientific journals.
• The number and coverage of systematic reviews is very limited. Available evidence is not properly integrated, in widespread industrial use, or of perceived value to stakeholders.	• Policy-makers, practitioners, and the general public have up-to-date and relevant systematic reviews and evidence-based guidelines and checklists at their disposal
• Lack of common terminology and appropriate descriptors and keywords, as well as limited retrieval facilities offered by electronic resources, hamper secondary SE research.	• The SE community is mature regarding the common understanding and use of basic terminology, descriptors and keywords. The electronic resources include all functionality that is needed to support the identification, selection, analysis and retrieval of bibliographic and full-text information about SE research.
• No standards for assessing the quality of primary and secondary research and thus of SE evidence.	• A common set of empirically-derived criteria for rating the quality of individual studies and for characterizing the overall strength of a body of evidence.
• No common understanding of SE phenomena.	• Agreed-upon conceptual and operational definitions of key SE constructs and variables.
• Limited advice on how to combine data from diverse study types.	• Methods are available for synthesizing evidence from a variety of perspectives and approaches to both research and practice.

4.4.2 Assessing the quality of primary studies. The usefulness of any systematic review depends on the quality of the primary studies available. Thus, quality assessment of primary studies that are included in systematic reviews is necessary to limit bias in conducting the review, gain insight into potential comparisons, and guide the interpretation of findings [61]. However, assessing the quality of primary SE studies poses a great challenge, and to the best of our knowledge, none of the existing systematic reviews in SE have included

quality assessments of primary studies as part of their criteria for inclusion.

The rating of the quality of experimental research has traditionally emphasized identifying threats to internal validity. However, the empirical basis for determining specific criteria for assessing quality for other empirical methods is less developed. Consequently, there are no standard methods for assessing the quality of data from qualitative, or mixed qualitative and quantitative, research. As the contribution of such research to the evidence base is increasingly acknowledged in SE, this poses a major challenge for performing high-quality systematic reviews.

There is also a debate about whether the concepts of quality used to assess qualitative research should be roughly the same as, parallel to, or quite different from, those used to assess quantitative research. There is also dispute about the extent to which quality assessment of qualitative inquiry can be formalized [53, 125].

There are also common problems in appraising the quality of published research, because journal articles and, in particular, conference papers rarely provide enough detail about the methods used, due to limitations of space in journal volumes and conference proceedings. Hence, there is a danger that what is being assessed is the quality of reporting, rather than the quality of research [59]. An important challenge is thus to provide, and critically examine, both empirically-derived and consensus-derived criteria for rating the quality of individual studies and for characterizing the overall strength of a body of evidence.

4.4.3 Assessing the scope of primary studies. Guidelines for empirical research in SE [77] recommend that authors define all interventions fully. However, several reviews of SE interventions have found limited descriptions of interventions and their underlying constructs [40, 56, 123]. As a consequence, there is little common understanding of important SE phenomena in terms of conceptual definitions and operationalized measures, which means that it is not possible to know what is actually meant by a specific construct in a particular study. It will thus be difficult (if not impossible) to compare the results of one study with those of another [38, 42].

Until SE researchers can agree upon the definitions of concepts used to describe the phenomena they study, we cannot go beyond face validity and ascertain accurately whether findings are comparable or not. A specific challenge is thus to define, conceptually and operationally, the constructs and variables used in empirical research and to explore the boundaries of construct and external validity.

4.4.4 Synthesizing results of heterogeneous studies. The key objective of research synthesis is to evaluate the included articles for heterogeneity and select appropriate methods for combining homogeneous studies [31]. The SE research literature often comprises heterogeneous interventions described in small studies that often differ significantly from each other with respect to the study design and outcomes evaluated. This heterogeneity often prevents quantitative meta-analysis of studies.

Since current procedures for systematic review are based on standard meta-analytic techniques, which are designed for combining data from homogeneous, quantitative studies [75], there is much less guidance on how to conduct reviews that incorporate qualitative and mixed-methods approaches.

Although research is underway in other disciplines (see, e.g., [36, 37]), there remain a number of methodological questions about the synthesis of qualitative findings. There are technical challenges, such as inter-rater reliability in abstracting qualitative data from individual studies and from intrastudy type syntheses to produce a cross-study type synthesis. There are also challenges related to the methods of qualitative synthesis, as well as to ways of integrating qualitative synthesis with meta-analysis. A key challenge, therefore, is to develop methods for synthesizing evidence from a variety of perspectives and approaches to research.

4.5 Theory building

In mature sciences, building theories is the principal method of acquiring and accumulating knowledge that may be used in a wide range of settings. There are many arguments in favor of using theories. They offer common conceptual frameworks that allow the organization and structuring of facts and knowledge in a concise and precise manner, thus facilitating the communication of ideas and knowledge. Furthermore, theory is the means through which one may generalize *analytically* [118, 134], thus enabling generalization from situations in which statistical generalization is not desirable or possible, such as from case studies [134] and across populations [91]. *Explanation* and *prediction* are important consequences of the above. Theory also helps common research agendas to be developed and consolidated (see Section 5.3).

Table 6. Source of theory

Source of Origin
Mode 1. Theories from other disciplines may be used as they are.
Mode 2. Theories from other disciplines may be adapted to SE before use.
Mode 3. Theories may be generated from scratch in SE.

Table 7. Sophistication of theory

Level of Sophistication
Level 1. Minor working relationships that are concrete and based directly on observations
Level 2. Theories of the middle range that involve some abstraction but are still closely linked to observations
Level 3. All-embracing theories that seek to explain SE behaviour. ("Social behavior" in [25] is here replaced with "SE")

Table 8. Use of theory

State of Practice	Target (2020-2025)
• Generally, little use of theories. The theories used mainly justify research questions and hypotheses; some explain results; very few test or modify theory.	• Most SE studies involve theories. Widespread use of theories entered in all three modes (Table 6). Considering using, testing, modifying or formulating theory is part of any empirical work
• Almost no SE-specific theories are proposed.	• Many SE theories are proposed and tested, and most of them are at Level 2 (Table 7) after having past the stage of Level 1. Level 3 seems still unrealistic.
• Theories are generally poorly documented	• There are widely used standards for describing theories in a clear and precise way.
• Difficult to identify the theories that actually are used or have been proposed.	• For each sub-discipline of SE, there are websites and systematic reviews that systematize and characterise relevant theories.

4.5.1 Use of theories. Although arguments in favor of proposing and testing theories based on empirical evidence in SE have been voiced in the SE community [7, 45, 60, 67, 77, 83, 115, 129], the use and building of such theories is still in its infancy. Nevertheless, some effort has been made. Hannay *et al*. [56] conducted a systematic review of the explicit use of theory in a comprehensive set of 103 articles reporting controlled experiments, from of a total of 5453 articles published in major SE journals and conferences in the decade 1993–2002. They found that of the 103 articles, 24 use a total of 40 theories in various ways to explain the cause-effect relationship(s) under investigation. The findings are summarized in Table 8.

Although the systematic review above was on controlled experiments, these findings are consistent with our experience from reading the literature, and being reviewers and members of program committees and editorial boards: theory use and awareness of theoretical issues are present in empirical studies, but empirically-based theory-driven research is, as yet, not a major issue in SE.

Table 6 shows three modes in which theories may be deployed to explain SE phenomena. Modes 1 and 2 reflect the fact that SE is multidisciplinary. Examples of Mode 1 are the use of theories from cognitive psychology to explain phenomena in program comprehension [1, 24, 109], and theories from social and behavioral sciences to explain group interaction in requirements negotiation and inspection meetings [83] Examples of Mode 2 can be found in [60, 83, 115], while Sjøberg *et al*. [122] give an example of Mode 3. At present, most SE theories are from other disciplines, possibly adapted; that is, they are theories of Modes 1 and 2. In the future, we hope to see greater use of theories in all modes. Greater use of theories specific to SE would require that they are actually built and tested; see below.

4.5.2 Building theories. Section 4.3 described the need for synthesis of empirical evidence. The next step is building theories. The term "theory" may be defined and evaluated in many ways [131]. What should constitute the nature of SE theories is an open issue. Nevertheless, since SE is an applied discipline, our position is that theories should have practical value for the software industry. Theories should provide input to general decision-making regarding technology and resource management. Theories should also help to explain and predict in given SE settings, but since each setting is unique, the theories would need local adaptations to be directly useful in a given setting.

Referencing [92, 135], Caroll and Swatman [25] describe theories in IS according to three levels of sophistication or complexity (Table 7). These levels set milestones in theory generation, but they may also rep-

resent full theories, depending on the rationale of the generation process to which one adheres and the purpose of one's theory. The development of SE theories from scratch is in the early stages, and immediate efforts will probably focus primarily on Levels 1 and 2. Level 1 theories will often have a narrow scope of application. Given the complexity of SE activities, the most useful theories will probably be at Level 2, with some context-specific information included in the theory (see point (c) below). Level 1 theories are candidates for being modified to become Level 2 theories. Another starting point for the formulation of Level 2 theories is SE principles, such as those collected by Glass [49] and Endres and Rombach [45].

To build more and better theories, the community needs to meet the challenges of Sections 4.1-4.4, that is, conducting many, high-quality, relevant empirical SE studies, including extensive combinations of studies in the form of replications [89] and families of studies [12], as well as synthesizing evidence from such studies.

4.5.3 Documenting theories. There is little guidance on how SE theories should be described and built. In the systematic review of theories [56], it was difficult to extract the theories from the text of the papers, because there are no uniformly accepted criteria for identifying theories and because little information was given about them. One important goal is to develop a common terminology for, and uniform way of describing, theories. In particular, the constructs, propositions and their explanations, and the scope of a theory, should be presented clearly and explicitly.

Some initial efforts have been made. Sjøberg *et al*. [122] propose that the constructs of an SE theory should be associated with one or more of the archetype classes shown in Table 1. Selecting or defining appropriate subclasses or component classes of these archetype classes also illustrates the need for commonly accepted taxonomies. If the constructs of SE theories do not follow from well-defined and well-understood categories of phenomena, new theories will frequently require new constructs, and then theories will become difficult to understand and to relate to each other.

4.5.4 Collecting theories. We have described above a systematic review of theories used in controlled experiments. In the future, there should be systematic reviews of the use of theories for other kinds of SE study as well. There should also be online resources for collecting and documenting theories in SE, following the lead in psychology (changing-minds.org/explanations/theories/theories.htm) and IS (www.istheory.yorku.ca/). Simula Research Laboratory has begun building a site for empirically-based SE theories. We believe that this will make it easier for scholars to find relevant theories for their research and that it will stimulate the community to collaborate on building new theories and on improving existing ones.

5. How to meet the challenges

This section discusses some ways in which the challenges described in the previous section might be met.

5.1 Competence: education and guidelines

The discrepancy between the state of practice of the application of empirical methods and the level at which we would like the research community to be (Section 4) indicates that there is a strong need to increase competence regarding how to conduct empirical studies and about the trade-offs among alternative empirical methods. This applies to researchers and reviewers, as well as to senior practitioners in industry. Although the quality of experiments in general should be much better in the future, the community does currently have a reasonable grasp of how to design, conduct, analyze and report experiments. However, understanding of other methods is generally low. For example, case studies are criticized on the grounds that it is not possible to generalize from them, which reflects a lack of understanding of the possibility of generalising analytically through theory. This, in turn, is related to a lack of understanding of what theory would mean in an SE setting; we must overcome the attitude that if SE theories do not take the same form as, for example, theories in physics, then we are not talking about theory [103].

One way of increasing the competence level is to integrate courses on empirical methods, synthesis of empirical evidence and theory building in SE education. As in many disciplines in the social and behavioral sciences, courses on these topics should be compulsory as part of any SE degree; see the discussion in [86, 108].

Providing appropriate guidelines for applying the various empirical methods is another means of increasing the competence level, particularly if the guidelines are based on systematic reviews of the state of practice or if they have been subject to empirical evaluation. So far, advice on, and guidelines for, the use of empirical methods in SE have emphasized experiments [40, 71, 77, 123, 133]. There are also guidelines for performing systematic reviews [75]. However, it still remains to produce text books, systematic reviews and guidelines that focus on case studies, action research, surveys and theory building that are tailored to particularly important challenges in SE contexts.

5.2 Collaboration between academia and industry

More and better collaboration between academia and the software industry is an important means of achieving the goals of more studies with high quality and relevance and better transfer of research results. A few examples follow. More empirical SE studies can be achieved by increasing research funding from industry and training software engineers in conducting more empirical studies with organization-specific goals. Studies of higher quality and greater relevance can be achieved by using the software industry as a laboratory, instead of studying inexperienced students performing small programming tasks. A more efficient transfer of research results may result from involving industry in the studies and rendering studies more credible by conducting them in contexts more similar to typical industrial contexts [121]. Nevertheless, collaboration with industry may sometimes lead to studies that have too many interfering factors for relationships to be understood or theories to be built, or to researchers being forced to focus on highly specific and short-term usefulness instead of more general and long-term results. To achieve the benefits while minimizing the risks, we have frequently found it useful, when collaborating with the software industry, to do the following:

- Draw a clear distinction between the research goals and the goals for improving processes of software development that are specific to the organisation [29]. This recommendation is based on our experience that although a software organization's work on improving processes of software development is quite similar to empirical research, its scope and the data collected may be too limited to allow high-quality studies. Hence, it is necessary to consider carefully the need for additional data collection or control to answer the research questions.
- Pay the software professionals ordinary fees for participating in studies designed by the researchers. We have found that this puts the industrial collaboration at ease, lowers the risk of study failure and increases the value of the study.
- Use opportunities at conferences, seminars and courses where software professionals meet to conduct SE studies. We have experienced that software professionals are more than willing to participate in small experiments (about 30 minutes' duration) on such occasions, particularly when they are informed that they will receive the results and learn from the feedback on the same day. Of course, not all types of research question can be addressed and there will be threats related to artificiality in such small experiments. If the experiments are combined with presentations on state-of-research within the topic addressed in the experiment, this is also an opportunity of research transfer.

In general, transferring results from academia to the software industry is difficult. It would be naïve to expect every researcher to have the ability and willingness to summarize and present research results that will be read by software professionals. Consequently, we believe that it is essential to adopt the following two measures to facilitate the transfer of research results:

- Engage more researchers with a particular focus on synthesizing results into SE guidelines and principles, in a language and format that make the results easy to understand and apply by the software industry.
- Educate science journalists who are not active researchers, but who have excellent writing skills, so that they develop a deep understanding of SE and research methods. The use of science journalists also has the advantage, compared with researchers summarizing their own results, that a more objective critique of the research is likely, because fewer vested interests are involved.

Further discussion on research collaborations between academia and industry can be found in [112].

5.3 Common research agendas

Addressing the challenges described in Section 4 will require more focus on empirical methods in the SE research community. The use of such methods should be solicited by journal editors, conference program chairs and reviewers. Moreover, to make significant progress in improving the quality of empirical studies, conducting more synthesis research, and building theories, there is a need for *common research agendas* or, in the terminology of Cohen [27], *cumulative research programs*. Many individual research groups undertake valuable empirical studies, but because the goal of such work is either individual publications and/or post-graduate theses, there is often a lack of overall purpose to such studies. Common research agendas would enable the researchers to ignore some issues as peripheral to the purpose of the research, and then to concentrate on central questions [27]. Replications and families of studies, including the use of different empirical methods, would be a natural part of such agendas.

Common research agendas should be established to improve empirical work per se, but also for specific SE topics, for example, distributed software development 117]. A more ambitious, long-term goal would be to

establish a program in SE similar to the Human Genome Project and CERN.

One challenge for such larger, joint efforts is to give credit to those who contribute and to ensure that data and study materials are dealt with properly. A template for agreements on how data and materials should be shared in a given setting can be found in [13].

5.4 Resources

Increasing the number of high-quality SE studies, primarily with the software industry as the laboratory, would require a substantial increase in the amount of resources available. To illustrate the kind of costs involved, consider a comprehensive experiment with professional developers. It may require several researchers and support staff for recruiting the subjects [18] and practical organization; funds for paying the professionals for their participation; and the development of infrastructure and experiment support apparatus. Hiring more than 100 consultants for one full day [4] may cost more than $100,000. A support tool that provides multiplatform support for downloading experimental materials and uploading task solutions, real-time monitoring of the experiment, recovery of experiment sessions, backup of experimental data, etc. was developed over several years and cost about $200,000 [5]. Case studies may also benefit from payment beyond salaries to researchers, for example, for data collection [2, 68].

In a given setting, the resources available will be limited. A researcher would have to prioritize between, for example, carrying out several artificial versus few comprehensive, realistic experiments, which is related to tradeoffs among internal, external, statistical conclusion and construct validity; see the discussion in [118]. Another challenge for the community at present is that prioritizing large, resource-intensive, longitudinal case studies or action research studies seems to lead to fewer publications than if the focus is on (say) small, simple experiments; a stronger emphasis on the former kind of studies must be reflected in the credit given by the community.

There are indications that more empirical work in SE is being funded. For example, there are an increasing number of members in the International Software Engineering Research Network (ISERN). However, pursuing our vision will require a significant increase in resources. In addition to funding more ordinary research positions, the SE community should ask for money for other purposes as well [23]; for example, for funding PhD students in industry [28] or supporting the conducting of empirical studies.

Finding the money to fund comprehensive empirical studies is a matter of politics. At Simula Research Laboratory, we have been given the flexibility to control our own budget in a way that we find optimal for our purposes, as long as we can envisage a good research outcome. Hence, we have decided to use about 25% of our budget for empirical studies, mainly at the expense of employing a larger number of researchers. Surprisingly, almost nobody in the community seems to include such expenses in their applications to funding bodies. The current attitude seems that empirical studies should be inexpensive, e.g., the use of students or inexpensive observation of industrial practice. In the future, applying for expenses for professionals or companies that take part in empirical studies should be as natural as applying for expensive laboratory hardware.

As SE researchers, we should contribute to making the development of software systems a mature industry. Given the importance of software systems in society [20], there is no reason why research projects in SE should be less comprehensive and cost less than large projects in other disciplines, such as physics and medicine. The U.S. funding for the Human Genome Project was $437 million over 16 years. If a wide range of scientific activities related to genomics are included, the total cost rises to $3 billion!

6. Summary

We presented a vision that the use of empirical methods could contribute to improved SE research and practice. Major challenges to reaching the vision were identified, i.e., more empirical studies of higher quality and relevance, and more focus on research synthesis and theory building. Each challenge was described in terms of the current state of the practice and how we envisage the targets for the future (2020-2025). Several ways of addressing these challenges were outlined, including increasing competence on conducting empirical studies, improving links between academia and industry, promoting common research agendas, and increasing resources devoted to empirical studies proportionate to the importance of software systems in society.

Acknowledgements

We thank Bente Anda, Erik Arisholm, Hans Christian Benestad, Gunnar Bergersen, James Dzidek, Stein Grimstad, Jo Hannay, Nina Elisabeth Holt, Vigdis By Kampenes, Amela Karahasanovic and Chris Wright for useful comments and discussions, and Chris Wright for proofreading.

References

[1] Abdel-Hamid, T.K., Sengupta, K. and Ronan, D. Software project control: an experimental investigation of judgement with fallible information, *IEEE Transactions on Software Engineering,* 19(6):603-612, 1993

[2] Anda, B.C.D., Benestad, H.C. and Hove, S.E., A Multiple-Case Study of Effort Estimation based on Use Case Points, In *ISESE'2005*. IEEE Computer Society, Noosa, Australia, Nov. 17-18, pp. 407–416, 2005

[3] Arisholm, E. and Sjøberg, D.I.K. Evaluating the Effect of a Delegated versus Centralized Control Style on the Maintainability of Object-Oriented Software, *IEEE Transactions on Software Engineering*, 30(8): 521-534, 2004

[4] Arisholm, E., Gallis, H.E., Dybå, T. and Sjøberg, D.I.K. Evaluating Pair Programming with Respect to System Complexity and Programmer Expertise, *IEEE Transactions on Software Engineering*, 33(2): 65-86, 2007

[5] Arisholm, E., Sjøberg, D.I.K., Carelius G.J. and Lindsjørn, Y. A Web-based Support Environment for Software Engineering Experiments, *Nordic Journal of Computing* 9(4): 231-247, 2002

[6] Avison, D., Lau, F., Myers, M. and Nielsen, P.A. Action Research, *Communications of the ACM*, 42(1): 94–97, 1999

[7] Basili, V.R. Editorial. *Empirical Software Engineering*, 1(2), 1996

[8] Basili, V.R. The Role of Experimentation in Software Engineering: Past, Current, and Future. 18th International Conference on Software Engineering (ICSE-18), Berlin, Germany, 25-29 March, pp. 442-449, 1996

[9] Basili, V.R., McGarry, F.E., Pajerski, R. and Zelkowitz, M.V. Lessons learned from 25 years of process improvement: the rise and fall of the NASA Software Engineering laboratory. ICSE-24, pp. 69-79, 2002

[10] Basili, V.R., Rombach, D., Schneider, K., Kitchenham, B., Pfahl, D. and Selby, R. *Experimental Software Engineering Issues: Assessment and Future*, Dagstuhl seminar, Germany, LNCS 4336, Springer-Verlag, 2007

[11] Basili, V.R., Selby, R.W. and Hutchens, D.H. Experimentation in Software Engineering, *IEEE Transactions on Software Engineering*, 12(7): 733–743, 1986

[12] Basili, V.R., Shull, F. and Lanubile, F. Building Knowledge through Families of Experiments, *IEEE Transactions on Software Engineering*, 25(4): 456–473, 1999

[13] Basili, V.R., Zelkowitz, M., Sjøberg, D.I.K., Johnson, P. and Cowling, T. Protocols in the Use of Empirical Software Engineering Artifacts, *Empirical Software Engineering*, 2007 (in press)

[14] Baskerville, R. and Wood-Harper, A.T. A Critical Perspective on Action Research as a Method for Information Systems Research, *Journal of Information Technology*, 11(3): 235–246, 1996

[15] Baskerville, R. and Wood-Harper, A.T. Diversity in Information Systems Action Research Methods, *European Journal of Information Systems*, 7(2): 90–107, 1998

[16] Benbasat, I. and Zmud, R.W., Empirical research in information systems: the practice of relevance, *MIS Quarterly*, 23(1): 3-16, 1999

[17] Benbasat, I., Goldstein, D.K. and Mead, M. The Case Research Strategy in Studies of Information Systems, *MIS Quarterly*, 11(3): 369–386, 1987

[18] Benestad, H.C., Arisholm, E. and Sjøberg, D.I.K. How to Recruit Professionals as Subjects in Software Engineering Experiments, IRIS'2005, 6-9 Aug., Kristiansand, Norway, 2005

[19] Boehm, B., Rombach, H.D. and Zelkowitz, M.V. (eds.) *Foundations of Empirical Software Engineering, The Legacy of Victor R. Basili*, Springer-Verlag, 2005

[20] Booch. G. Developing the future, *Communications of the ACM*, (44(3): 118-121, 2001

[21] Brehmer, B. In One Word: Not from Experience, *Acta Psychologica*, 45(1-3): 223–241, 1980

[22] Briand L. and Wolf A. (eds.) *Future of Software Engineering,* IEEE-CS Press, 2007

[23] Brilliant, S.S. and Knight, J.C., Empirical Research in Software Engineering: A Workshop, *Software Engineering Notes,* 24(3): 45-52, 1999

[24] Burkhardt, J.M., Detienne, F. and Wiedenbeck, S. Object-Oriented Program Comprehension: Effect of Expertise, Task and Phase, *Empirical Software Engineering*, 7(2): 115-156, 2002

[25] Carroll, J. and Swatman P.A. Structured-Case: A methodological framework for building theory in IS research, European Conference on IS, Vienna, 3-5, July 2000

[26] Christensen, L.B. *Experimental Methodology*, 10th ed. Allyn & Bacon, 2006

[27] Cohen, B.P. Developing Sociological Knowledge: Theory and Method, 2nd Ed., Chicago: Nelson-Hall, 1989

[28] Conradi, R., Dybå, T. Sjøberg, D.I.K. and Ulsund, T. *Software Process Improvement: Results and Experience from the Field*, Springer-Verlag, 2006

[29] Conradi, R., Dybå, T. Sjøberg, D.I.K. and Ulsund, T., Lessons Learned and Recommendations from two Large Norwegian SPI Programmes, EWSPT'2003, Oquendo (ed.), Helsinki, Finland, Springer-Verlag, pp. 32-45, 2003

[30] Cook, T.D. and Campbell, D.T. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company, 1979

[31] Cooper, H. and Hedges, L.V. (eds.) *Handbook of Research Synthesis*, NY: Russell Sage Foundation, 1994

[32] Cooper, H. *Synthesizing Research* (3rd Ed.), Thousand Oaks, CA: Sage, 1998

[33] Dahl, O.J. and Nygaard, K. SIMULA: An ALGOL-based Simulation Language, *Communications of the ACM*, 9(9): 671-678, 1966

[34] Davison R.M., Martinsons, M.G. and Kock, N. Principles of Canonical Action Research, *Information Systems Journal*, 14: 65–86, 2004

[35] DeVellis, R.F. *Scale Development: Theory and Applications* (2nd Ed.), Newbury Park, Ca: Sage, 2003

[36] Dixon-Woods, M., Agarwal, S., Jones, D., Young, B. and Sutton, A. Synthesising Qualitative and Quantitative Evidence: A Review of Possible Methods, *Journal of Health Services Research & Policy*, 10(1): 45–53, 2005

[37] Dixon-Woods, M., Bonas, S., Booth, A., Jones, D.R., Miller, T., Sutton, A.J., Shaw, R.L., Smith, J.A. and Young, B. How can systematic reviews incorporate qualitative research? A critical perspective, *Qualitative Research*, 6(1): 27–44, 2006

[38] Dybå, T. An Empirical Investigation of the Key Factors for Success in Software Process Improvement, *IEEE Transactions on Software Engineering*, 31(5): 410–424, 2005

[39] Dybå, T. An Instrument for Measuring the Key Factors of Success in Software Process Improvement, *Empirical Software Engineering*, 5(4): 357–390, 2000

[40] Dybå, T., Kampenes, V.B. and Sjøberg, D.I.K. A Systematic Review of Statistical Power in Software Engineering Experiments, *Information and Software Technology*, 48(8): 745–755, 2006

[41] Dybå, T., Kitchenham, B.A. and Jørgensen, M. Evidence-Based Software Engineering for Practitioners, *IEEE Software*, 22(1): 58–65, 2005

[42] Dybå, T., Moe, N.B. and Arisholm, E. Measuring Software Methodology Usage: Challenges of Conceptualization and Operationalization, *ISESE'2005*, Noosa Heads, Australia, 17-18 Nov., pp. 447-457, 2005

[43] Egger, M., Smith, G.D. and Altman, D.G. *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd Ed.), London: BMJ Publishing Group, 2001

[44] Elden, M. and Chisholm, R.F. Emerging Varieties of Action Research: Introduction to the Special Issue, *Human Relations*, 46(2): 121–142, 1993

[45] Endres, A. and Rombach, D. *A Handbook of Software and Systems Engineering. Empirical Observations, Laws and Theories*. Fraunhofer IESE Series on Software Engineering. Pearson Education Limited, 2003

[46] Fenton, N., Pfleeger, S.L. and Glass, R.L. Science and Substance: A Challenge to Software Engineers, *IEEE Software*, 11(4): 86-95, 1994

[47] Fink, A. and Kosecoff, J. *How to Conduct Surveys: A Step-By-Step Guide* (3rd Ed.), Thousand Oaks, California: Sage Publications, 2005

[48] Gerring, J. *Case Study Research: Principles and Practices*, New York: Cambridge Univ. Press, 2007

[49] Glass, R.L. *Facts and Fallacies of Software Engineering*, Addison-Wesley, 2003

[50] Glass, R.L. Pilot Studies: What, Why and How, *Journal of Systems and Software*, 36(1): 85–97, 1997

[51] Glass, R.L., Vessey, I. and Ramesh, V. Research in Software Engineering: An Analysis of the Literature, *Information and Software Technology*, 44(8): 491-506, 2002

[52] Glass, R.L. and Vessey, I. Contemporary Application-Domain Taxonomies. *IEEE Software,* 12(4): 63-76, 1995

[53] Greenhalgh, T. *How to Read a Paper* (3rd Ed.), London: BMJ Publishing Group, 2006

[54] Greenwood, D.J. and Levin, M. *Introduction to Action Research: Social Research for Social Change* (2nd Ed.), Thousand Oaks, California: Sage, 2006

[55] Haig, B.D. An Abductive Theory of Scientific Method, *Psychological Methods*, 10(4): 317–388, 2005

[56] Hannay, J., Sjøberg, D.I.K. and Dybå, T. A Systematic Review of Theory Use in Software Engineering Experiments, *IEEE Transactions on Software Engineering*, 33(2): 87-107, 2007

[57] Hardy, M. and Bryman, A. *Handbook of Data Analysis*, London: Sage Publications, 2004

[58] Harrison, W. Skinner Wasn't a Software Engineer, Editorial, *IEEE Software*, 22(3): 5-7, 2005

[59] Hawker, S., Payne, S., Kerr, C., Hardey, M. and Powell, J. Appraising the Evidence: Reviewing Disparate Data Systematically, *Qualitative Health Research*, 12(9): 1284–1299, 2002

[60] Herbsleb, D.J. and Mockus, A. Formulation and Preliminary Test of an Empirical Theory of Coordination in Software Engineering, *ACM SIGSOFT Software Engineering Notes*, 28(5): 138-147, 2003

[61] Higgins J.P.T. and Green S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6* [September 2006]. In The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd. 2006

[62] Höfer, A. and Tichy, W.F. Status of Empirical Research in Software Engineering. In: Basili *et al.* (eds), Experimental Software Engineering Issues: Assessment and Future Directions, Springer-Verlag, LNCS 4336, 2007

[63] Holt, N.E. A Systematic Review of Case Studies in Software Engineering, MSc thesis, Dep. of Informatics, Univ. of Oslo, 2006

[64] Hove, S.E. and Anda, B.C.D. Experiences from Conducting Semi-Structured Interviews in Empirical Software Engineering Research, METRICS'2005, 19-22 Sep., Como, Italy. IEEE, pp. 1-10, 2005

[65] Jørgensen, M. A Review of Studies on Expert Estimation of Software Development Effort, *Journal of Systems and Software*, 70(1-2): 37–60, 2004

[66] Jørgensen, M. and Shepperd, M. A Systematic Review of Software Development Cost Estimation Studies, *IEEE Transactions on Software Engineering*, 33(1): 33–53, 2007

[67] Jørgensen, M. and Sjøberg, D.I.K. Generalization and Theory-Building in Software Engineering Research, EASE'2004, *IEE Proceedings*, pp. 29–36, 2004

[68] Jørgensen, M. Realism in Assessment of Effort Estimation Uncertainty: It Matters How You Ask. *IEEE Transactions on Software Engineering*, 30(4): 209-217, 2004

[69] Jørgensen, M. Software Quality Measurement, *Advances in Engineering Software,* 30(12): 907-912, 1999

[70] Jørgensen, M., Dybå, T. and Kitchenham, B.A. Teaching Evidence-Based Software Engineering to University Students, *Proceedings of METRICS'2005*, Como, Italy, 19-22 Sept. 2005

[71] Juristo, N. and Moreno, A.M. *Basics of Software Engineering Experimentation*, Dordrecht, Kluwer Academic Publishers, 2001

[72] Kahneman, D, and Tversky, A. (eds.), *Choices, Values and Frames*. New York: Cambridge University Press and the Russell Sage Foundation, 2000

[73] Kampenes, V.B., Dybå, T., Hannay, J.E. and Sjøberg, D.I.K. A Systematic Review of Effect Size in Software Engineering Experiments, to appear in *Journal of Information and Software Technology*, 2007

[74] Kitchenham, B.A. Evaluating Software Engineering Methods and Tools, Parts 1 to 12, *SIGSOFT Software Engineering Notes*, Vols. 21-23, 1996/1998

[75] Kitchenham, B.A. Procedures for Performing Systematic Reviews, Keele University, Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1. 2004

[76] Kitchenham, B.A., Dybå, T. and Jørgensen, M. Evidence-Based Software Engineering, ICSE'2004, Edinburgh, Scotland, 23-28 May, pp. 273–281, 2004

[77] Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K. and Rosenberg, J. Preliminary Guidelines for Empirical Research in Software Engineering, *IEEE Transactions on Software Engineering*, 28(8): 721–734, 2002

[78] Kitchenham, B.A., Pickard, L.M. and Pfleeger, S.L. Case Studies for Method and Tool Evaluation, *IEEE Software*, 12(4):52–62, 1995

[79] Klein, H.K. and Myers, M.D. A Set of Principles for Conducting and Evaluating Interpretive Field Studies in IS, *MIS Quarterly*, 23(1): 67–93, 1999

[80] Kock, N. (ed.) *Information Systems Action Research: An Applied View of Emerging Concepts and Methods*, Berlin: Springer-Verlag, 2007

[81] Kraut, A.I. (ed.) *Organizational Surveys: Tools for Assessment and Change*, San Francisco: Jossey-Bass, 1996

[82] Krosnick, J.A. Survey Research, *Annual Review of Psychology*, 50: 537–567. 1999

[83] Land, L.P.W., Wong, B. and Jeffery, R. An Extension of the Behavioral Theory of Group Performance in Software Development Technical Reviews, Tenth Asia-Pacific SE Conference, 520-530, 2003

[84] Lau, F. Toward a Framework for Action Research in Information Systems Studies, *Information Technology & People*, 12(2): 148–175, 1999

[85] Lee, A.S. A Scientific Methodology for MIS Case Studies, *MIS Quarterly*, 13(1): 33–50, 1989

[86] Lethbridge, T.C., Daz-Herrera, J., LeBlanc Jr., R.J., Thompson, J.B. Improving Software Practice through Education: Challenges and Future Trends. In [22]

[87] Lethbridge, T.C., Sim, S.E. and Singer, J. Studying Software Engineers: Data Collection Techniques for Software Field Studies, *Empirical Software Engineering*, 10(3): 311–341, 2005

[88] Lindblom, C.E. Alternatives to Validity. Some Thoughts Suggested by Campbell's Guidelines. *Knowledge: Creation, Diffusion, Utilization*, 8: 509-520, 1987

[89] Lindsay, R.M. and Ehrenberg, A.S.C. The Design of Replicated Studies, *The American Statistician*, 47(3): 217–228, 1993

[90] Lipsey, M.W. and Wilson, D.B. *Practical Meta-Analysis*, Thousand Oaks, CA: Sage, 2001

[91] Lucas, J.W. Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, 21: 236–253, 2003

[92] Merton, R.K. *Social theory and social structure*, 3rd ed., The Free Press, New York, 1968

[93] Miles M.B. and Huberman, A.M. *Qualitative Data Analysis: An Expanded Source Book*, Thousand Oaks, CA: Sage, 1994

[94] Miller, J. Applying Meta-analytical Procedures to Software Engineering Experiments, *Journal of Systems and Software*, 54(1): 29–39, 2000

[95] Montgomery, D.C. *Design and Analysis of Experiments* (6th Ed.), New Jersey: John Wiley & Sons, 2004

[96] Mulrow, C. and Cook, D. (eds.) *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions*, Philadelphia, PA: Am. College of Physician, 1998

[97] Neff, F.W. Survey Research: A Tool for Problem Diagnosis and Improvement in Organizations. In A.W. Gouldner and S.M. Miller (eds.), *Applied Sociology*, New York: Free Press, pp. 23-38, 1966

[98] Neuman, W.L. *Social Research Methods: Quantitative and Qualitative Approaches* (6th Ed.), Boston: Allyn & Bacon, 2006

[99] Noblit, G.W. and Hare, R.D. *Meta-Ethnography: Synthesizing Qualitative Studies*, Sage University series on Qualitative Research Methods, Vol. 11, Thousand Oaks, CA: Sage, 1988

[100] Paterson, B.L., Thorne, S.E., Canam, C., and Jillings, C. *Meta-Study of Qualitative Health Research: A Practical Guide to Meta-Analysis and Meta-Synthesis*, Thousand Oaks: Sage., 2001

[101] Perry, D.E., Porter, A.A. and Votta, L.G. Empirical studies of Software Engineering: a roadmap, ICSE – Future of Software Engineering Track, pp. 345-355, 2000

[102] Petticrew, M. and Roberts, H. *Systematic Reviews in the Social Sciences: A Practical Guide*, Oxford, UK: Blackwell, 2006

[103] Pfleeger, S.L. Albert Einstein and Empirical Software Engineering, *IEEE Computer*, 32(10): 32–38, Oct. 1999

[104] Pfleeger, S.L. Experimental Design and Analysis in Software Engineering, Parts 1 to 5, *SIGSOFT Software Engineering Notes*, Vols. 19-20, 1994/1995

[105] Pickard, L.M., Kitchenham, B.A., Jones, P.W. Combining empirical results in Software Engineering, *Information and Software Technology*, 40(14): 811–821, 1998

[106] Pinsonneault, A. and Kraemer, K.L. Survey Research Methodology in Management Information Systems: An Assessment, *Journal of Management Information Systems*, 10(2): 75–105, 1993

[107] Popper, K. *The Logic of Scientific Discovery*, Hutchison, London, 1959

[108] Rainer, A., Ciolkowski, M., Pfahl, D., Kitchenham, B., Morasca, S. and Müller, M.M., Travassos, G.H., Vegas, S. Teaching Empirical Methods to Undergraduate Students. In [10], pp. 158–162

[109] Ramanujan, S., Scamell, R.W. and Shah, J.R. An Experimental Investigation of the Impact of Individual, Program, and Organizational Characteristics on Software Maintenance Effort, *Journal of Systems and Software*, 54(2): 137-157, 2000

[110] Reason, P. and Bradbury, H. (eds.) *Handbook of Action Research: Participative Inquiry and Practice*, Thousand Oaks, California: Sage, 2001

[111] Robson, C. *Real World Research* (2nd Ed.), Oxford: Blackwell Publishers, 2002

[112] Rombach D. and Achatz, R. Research Collaborations between Industry and Academia in a Global World. In [22]

[113] Rombach, D.H., Basili, V.R. and Selby, R.W. (eds.): Experimental Software Engineering Issues: Critical Assessment and Future Directions, Int. Workshop, Dagstuhl seminar, Germany, Sep. 14-18, 1992, LNCS 706 Springer-Verlag, 1993

[114] Rosenthal, R. and DiMatteo, M.R. Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews, *Annual Review of Psychology*, 52: 59-82, 2001

[115] Sauer, C, Jeffery, D.R., Land, L. and Yetton, P. The Effectiveness of Software Development Technical Reviews: A Behaviorally Motivated Program of Research, *IEEE Transactions on Software Engineering*, 26(1): 1-14, 2000

[116] Seaman, C. Qualitative Methods in Empirical Studies of Software Engineering, *IEEE Transactions on Software Engineering*, 25(4): 557-572, 1999

[117] Sengupta, B., Chandra, S. and Sinha, V. A Research Agenda for Distributed Software Development, ICSE'2006, Beijing, pp. 731-740, 2006

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

[119] Sjøberg, D. Use of Professionals in Experiments, Letter to Editor, *IEEE Software*, 22(5): 9-10, 2005

[120] Sjøberg, D.I.K., Anda, B., Arisholm, E., Dybå, T., Jørgensen, M., Karahasanovic, A., Koren, E.F. and Vokac M. Conducting Realistic Experiments in Software Engineering, *Proceedings ISESE'2002*, Nara, Japan, 3-4 October, pp. 17-26, 2002

[121] Sjøberg, D.I.K., Anda, B., Arisholm, E., Dybå, T., Jørgensen, M., Karahasanovic, A. and Vokac M. Challenges and Recommendations when Increasing the Realism of Controlled Software Engineering Experiments. In: Conradi and Wang (eds.) *Empirical Methods and Studies in Software Engineering: Experiences from ESERNET*, Berlin: Springer-Verlag LNCS 2765, pp. 24-38, 2003

[122] Sjøberg, D.I.K., Dybå, T. Anda, B.C.D. and Hannay, J.E. Building Theories in Software Engineering. To appear in F. Shull, J. Singer, D.I.K. Sjøberg (eds.), *Advanced Topics in Empirical Software Engineering*, Springer-Verlag, 2007

[123] Sjøberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanović, A., Liborg, N.-K. and Rekdal, A.C. A Survey of Controlled Experiments in Software Engineering, *IEEE Transactions on Software Engineering*, 31(9): 733–753, 2005

[124] Spector, P. *Summated Rating Scale Construction: An Introduction*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-082, Newbury Park, California: Sage, 1992

[125] Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*, London: Government Chief Social Researcher's Office, 2003

[126] Stake, R.E. *The Art of Case Study Research*, Thousand Oaks, California: Sage, 1995

[127] Thorne, S., Jensen, L., Kearney, M.H., Noblit, G. and Sandelowski, M. Qualitative Metasynthesis: Reflections on Methodological Orientation and Ideological Agenda, *Qualitative Health Research*, 14(10): 1342–1365, 2004

[128] Tichy, W.F. Hints for Reviewing Empirical Work in Software Engineering, *Empirical Software Engineering*, 5(4): 309-312, 2000

[129] Tichy, W.F. Should computer scientist experiment more? 16 excuses to avoid experimentation. *IEEE Computer*, 31(5): 32–40, 1998

[130] Tichy, W.F., Lukowicz, P., Prechelt, L. and Heinz, E.A. Experimental Evaluation in Computer Science: A Quantitative Study, *Journal of Systems and Software*, 28(1): 9-18, 1995

[131] Torraco, R.J. and Holton, E.F. A Theorist's Toolbox, *Human Resource Development Review*, 1(1): 129-140, 2002

[132] Weick, K.E. Theory Construction as Disciplined Imagination, *Academy of Management Review*, 14(4): 516-531, 1989

[133] Wohlin, C., Runeson, P. Höst, M., Ohlsson, M.C., Regnell, B. and Wesslén, A. *Experimentation in Software Engineering: An Introduction*, Boston: Kluwer Academic Publishers, 2000

[134] Yin, R.K. *Case Study Research: Design and Methods* (3rd ed.), Thousand Oaks, CA: Sage, 2003

[135] Yin, R.K., *Case Study Research: Design and Methods*, Sage Publications, CA., 1984

[136] Zelkowitz, M.V. and Wallace, D. Experimental Models for Validating Technology, *Theory and Practice of Object Systems*, 31(5): 23-31, 1998

[137] Zelkowitz, M.V. and Wallace, D. Experimental Validation in Software Engineering, *Information and Software Technology*, (39): 735-743, 1997

[138] Zeller, A. The Future of Programming Environments: Integration, Synergy, and Assistance. In: [22]

[139] Zendler, A. A Preliminary Software Engineering Theory as Investigated by Published Experiments, *Empirical Software Engineering*, 6(2): 161–180, 2001