Empirical Practice in Software Engineering

Andreas Jedlitschka, Liliana Guzmán, Jessica Jung, Constanza Lampasona, Silke Steinbach

Fraunhofer Institute for Experimental Software Engineering,
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
{surname.name@iese.fraunhofer.de}

Abstract Experimental software engineering was defined as the scientific approach to systematically evaluate software technologies referring to predefined hypotheses using sound empirical methods.
The purpose of this chapter is to give an overview of the history, current practice, and future of empirical practice in Software Engineering. In particular, based on what we learned from 20 years of research in empirical software engineering, we describe our currently used empirical approach in terms of a scientific approach to applied research and as a means for systematic evaluation.

1. Introduction

The Origins of Experimentation in Software Engineering

Since the very beginning, empirical Software Engineering (SE) was thought to be a rigorous discipline that should help industrial decision makers. In 1986, Basili published his fundamental paper on Experimentation in SE and stated [BSH86]: *"… experimentation is performed to help us better evaluate, predict, understand, control and improve the software development process and product"*. Wohlin et al. defines: [WR+00]: *"Experimentation provides a systematic, disciplined, quantifiable and controlled way of evaluating human-based activities"*. Bringing Basili's statement and Wohlin's et al. definition together with the IEEE definition of SE, the potential of experimentation becomes clear, especially with regard to a systematic and quantifiable approach [IEEE90]: *"Software Engineering means application of a systematic, disciplined, quantifiable approach to the development, operation and maintenance of software."* Juristo and Moreno reformulate Basili's claim in the following way [JM01]: *"The aim of SE experimentation is to provide facts for the suppositions, assumptions, speculations and beliefs that abound in software construction"*.

This means, experimentation is the scientific approach to systematically evaluate software technologies referring to predefined hypotheses using sound empirical methods.

Therefore, experimentation is considered a major concept in organizational learning and an important means for supporting decision making at the management level, especially in case of software process improvement (SPI). It is widely accepted that experimentation provides deeper insights into understanding and establishing cause-effect relationships than any other evaluation method [Sch70].

Establishing Empirical Software Engineering

The first Dagstuhl Seminar on Experimental SE was in 1992 [RBS92]. Researchers from both SE and experimentation in SE came together for discussing the state of the art and practice of experimentation in SE and for proposing a future research agenda. They concluded that the empirical methods applied in SE were mostly restricted to quantitative studies, i.e., controlled experiments. Since then a range of qualitative studies has been introduced, from observational to ethnographical studies. Thus the field moved from experimental to empirical SE.

In 1996, D. Rombach founded the Fraunhofer Institute for Experimental Software Engineering [Fraunhofer IESE]. Over time Fraunhofer IESE has become one of the most important research institutes in empirical SE, contributing with multiple empirical studies to the knowledge base of SE. Additionally it has contributed to educate SE researchers in empiricism, encourage software organizations to get involved in empirical studies, and adapt empirical methods the SE context [Rom00].

In 1993 the International Software Engineering Research Network (ISERN) [ISERN] was initiated by D. Rombach, V. Basili, R. Jeffery, G. Cantone, M. Oivo, and K. Torii. Since then ISERN grew to about 60 members (2012) around the world. ISERN members commit themselves to use the experimental paradigm in their research. The aim of ISERN was supported by many authors. Fenton et al. [FPG94] claimed repeatedly that rigorous experimentation is needed to evaluate new software technologies and their effects on organizations, processes, and products. Moreover, Pfleeger stated that [Pfl94]: *"As a software manager, it is important to make key decisions or assessments in an objective and scientific way"*.

After the first Dagstuhl Seminar on Experimental SE and the foundation of ISERN, several (some of them independent) initiatives have been performed for increasing the acceptance and use of empirical methods in SE.

In the context of the U.K. DESMET project [KLL97] Kitchenham and colleagues developed a methodology for evaluating SE technologies by using empirical methods. The commonly used/published types of evaluations are post-mortems, surveys, case studies, and both controlled and quasi experiments.[1]

The growing interest in empirical methods also required rigor in performing studies as well as in reporting their results. Consequently the discussion about guidelines for the empirical work in SE started. In 1999, Singer [Sin99] described

---

[1] In this chapter, we use the word study for all kinds of empirical studies. We use experiment as the generic term for controlled and quasi-experiments.

how to use the "American Psychological Association (APA) Styleguide" [APA01] for publishing experimental results in SE. In parallel, the first text books regarding experimentation in SE were published starting in 2000.

The first text books regarding experimentation in software engineering were published starting in 2000. Wohlin et al. [WR+00] provide an introduction to experimentation with the focus on the evaluation of methods, techniques, and tools in software engineering. They describe the whole experimental process starting with the definition and ending with the presentation. Juristo and Moreno [JM01] provide the basics for experimentation in software engineering, especially for planning, conducting, and analyzing experiments. Because surveys were at that time among the research methods employed most, Pfleeger and Kitchenham started a series on the principles of survey research [PK01].

Also in 2001, a first version of the initial guidelines on how to perform, report, and collate results of empirical studies in SE based on medical guidelines as well as on the personal experience of the authors was published [KP+02].

In 2003, Shaw [Sha03] provided a tutorial on how to write scientific papers, including the presentation of empirical research as a special case.

Maturing Empirical Software Engineering

With the increase in maturity and understanding of empirical SE, further requirements were formulated. For instance, in 2003 Ruhe [Ruh03] argued that empirical SE has to prove its industrial value by its contribution to decision support.

Moreover several attempts were initiated to build a common body of knowledge based on the synthesis of existing empirical evidence. Until then, empirical SE research yielded a considerably large number of scientific publications ranging from controlled experiments and countless numbers of case studies and surveys. However, attempts to combine the knowledge gained in single studies is reported to be extremely effort consuming [APW02, JMV05], if not impossible [JC04]. Consequently, a body of knowledge was built only for a few areas of SE (i.e., inspections and testing). Besides the individual efforts to build a common body [APW02, WPA03, JMV05], projects like the European ESERNET [CW03], the American CeBASE, and the Norwegian SPI Programmes [CD+03] aimed at providing such a kind of knowledge.

However, at that time, there was clearly no concise answer to R. Glass' [Gla04] request for software managers' support from research. Glass, taking the standpoint of software managers, wrote: *"Here's a message from software managers to software researchers: We (software managers) need your help. We need some better advice on how and when to use methodologies"*.

At the same time, Turner [Tur04], acknowledging the difficulties in really responding to these demands, stated that empiricism (if applied in a goal-oriented manner and not for the pure sake of quantification) can help to answer the following important questions regarding the "value" of a technology: what are the real costs, what is the benefit, what is its origin, in which context it can be applied, what is the latency, and what might be the barriers.

Also in 2004, a new trend, namely evidence-based SE, came up, which was mainly from evidence-based medicine. The evidence-based paradigm "… *proposes the use of currently best empirical evidence from research integrated with practical experience and human value judgment to support decision-making processes in the development and maintenance of software.*" [KDJ04].

Based on the evidence-based paradigm, Kitchenham [Kit04] proposed a guideline for conducting and reporting systematic reviews.

Following on that, researcher in empirical SE investigate to which extent the measures implemented in evidence-based medicine can be transferred to and adopted by SE. One issue arises in SE, for example, from the difference in the numbers of publications, especially on controlled experiments. A systematic literature of experiments in SE, [SH+05] identified 103 controlled experiments for the years 1993 to 2002, however, for medicine, 97,467 randomized controlled trials were published for the same period in PubMED. Despite the availability of text books and guidelines, Sjøberg et al. [SH+05] confirm our earlier findings [JC04] that reporting of results from studies still is often vague, unsystematic, and lacking consistent terminology. In their conclusion, Sjøberg et al. recommend that researchers should accurately report "… *the type and number of subjects, including the mortality rate; context variables such as general software engineering experience and experience specific to the tasks of the experiments; how the subjects were recruited; the application areas and type of tasks; the duration of the tasks; and internal and external validity of the experiments, including being specific about the sample and target population of the experiment.*"

Standards for reporting results were asked for to facilitate the review of articles, ease replication of experiments and any kind of synthesis as well as theory building. Based on existing guidelines, requirements from the field and feedback from the community, we iteratively developed a guideline for reporting results from controlled experiments [JP05].

The 2006 Dagstuhl Seminar on Empirical SE [BR+07] yielded that since 1992, the topic of empirical SE has been adopted more widely by academia as an interesting and promising research topic, and by industry practice as a necessary infrastructure technology for goal-oriented, sustained process improvement. At the same time, the spectrum of methods applied in empirical SE has broadened. Participants acknowledged empirical SE a positive evolution. However, open issues were identified, among them, the need for better support of the reuse of empirical knowledge (combination of results) and further standardizing the way, on how empirical studies are performed and reported.

Since then, the number of (systematic) reviews significantly increased, addressing several topics including, e.g., elicitation techniques [DJ11], and agile development [DD08]. However, researcher often used narrative summary for synthesizing[2] the results from individual studies [Cio09].

In addition, guidelines for reporting experiments [JCP08], case studies [RH09], replications [Car10] were further discussed and consolidated.

Current research on empirical methods in SE is dedicated to understand and extend existing quantitative [Cio09, Cio12] and qualitative synthesis [CD11] methods for better addressing the specific needs in SE, and how to derive laws and theories from empirical evidence [Rom11].

2. Empirical research process at Fraunhofer IESE

In 1997, D. Rombach described an "*experimentally-based software technology transfer*" concept based on the Goal/Question/Metric (GQM) method [BCR01b], the Quality Improvement Paradigm (QIP) [BCR01a], and the Experience Factory (EF) [BCR01a].

In 2013, applied research and technology transfer at Fraunhofer IESE is enriched by an evolved understanding of a goal-oriented research process, e.g., of what has to be done and what alternatives are available at the different stages of the QIP.

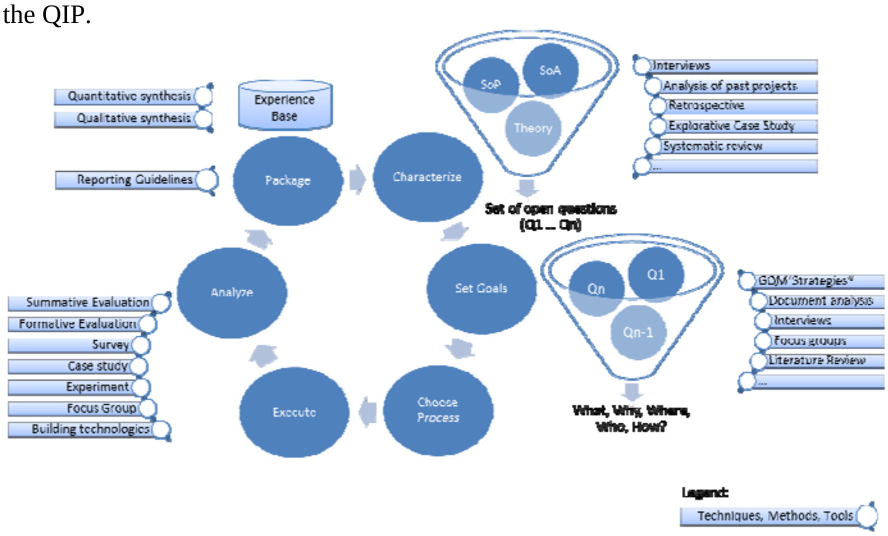

Figure 1. The Empirical Research Process

In the following subsections, we present the Fraunhofer IESE applied research process, which is defined on the basis of the QIP. It includes the following steps (cf. Figure 1):

[2] Synthesis is the umbrella term covering different strategies for combining empirical evidence [CHC02].

1. Characterize: Characterize the problem, i.e., identify, and specify what the problem is. Experience from previous projects is incorporated.
2. Set goals: Define the research goal, i.e., what, why, and how. This comprises the selection of appropriate solution strategies for achieving the predefined goals.
3. Choose Process: Choose and describe the process to implement the strategies, i.e., establish a plan how the research will be conducted, in particular, what will happen during the execution, incl. the design of empirical studies.
4. Execute: Perform the research according to the plan, i.e., build technologies or models, introduce them into the organization, collect data regarding goal achievement and enhancing models.
5. Analyze: Summative analysis of results and evaluate the degree of goal achievement.
6. Package: Document, report and disseminate results of all previous phases. If appropriate, integrate the results with the existing body of knowledge.

The steps of the process support us focusing our research on problems that are relevant for SE practice and thus support our mission of applied research. The process is applicable for all kinds of projects, e.g., customer projects, research projects, and individual projects. Furthermore, each step is enriched by empirical studies and methods. The theory of empirical research processes and methods can be found in several test books, such as [Cre03, Yin03, SCC02, WR+00].

Characterize

For both customer and research projects, the first activity is to profoundly formulate the problem of interest, in particular, the what, why, and how [CR83]. While customer projects are driven by more or less explicitly defined organizational goals and context thus providing well defined starting points at different stages of the process, research projects have a large degree of freedom for specifying the scope of interest. Nevertheless the general process remains the same.

According to our process, the first step is to identify and characterize problems in SE practice, e.g., by interviewing or surveying experts in the field, e.g., [CLB03, JH+05, JC+07], analyzing past projects by performing, e.g., data analysis [KN+10], project retrospectives, or strengths-weakness analysis.

For the retrospectives, which we successfully used in projects with several customers, we adapted the approach proposed by Kerth [Ker01]. The scope varies from single-shot identification of improvement potential from projects to long-term organizational learning. The final results of the retrospectives provide valuable feedback to our customers (e.g., identification of improvement potential) and important input for our research agenda. Our approach for performing retrospective is supported by empirical methods including, e.g., small-sized surveys, document analysis, and focused workshops.

To analyze the current situation at an organization, we employ strengths and weaknesses analysis based on maturity and capability reference models enriched with an analysis of work-products. Also here we gain a rich insight into the state of the practice, even if it is restricted to the specific context of the organization. Our experiences have shown that semi-structured interviews and document analysis are effective empirical methods for supporting this kind of analysis [JH+05].

If applicable, we additionally review the state of the art in SE by performing systematic reviews. Here we usually focus on: (1) identifying existing solutions, i.e., software technologies, (2) synthesizing empirical evidence regarding the effects of those software technologies in the context of interest (i.e., identifying existing theories), and (3) identifying questions that research does yet not have an answer for. Especially for research projects this is an important task, because the research has to be linked to existing theories. An introduction into the topic of theories in SE is given, e.g., in [SD+08].

To get a deeper insight into the problem and its context, we perform quantitative or qualitative studies. In the first case, we use, e.g., surveys, explorative pilot studies, or available data, and in the latter, we use structured questionnaires, observational studies, or focus groups.

At this point we have identified a set of open questions (i.e., problems to be solved).

Set goal

Together with the experts, we prioritize the open questions (i.e., problems) and define the scope (i.e., number of problems addressed).

Then we translate the problems into related goals, and elicit first improvement suggestions (this is typically part of our retrospectives). Within the target scope, problems are translated into initial goals. Then goals are further refined, prioritized and specified. Afterwards appropriate strategies are chosen.

In general, we derive appropriate goals and related strategies together with rationales by using the GQM^+Strategies® approach [BH10]. With the GQM^+Strategies® approach, we ensure the relevance and coherence of goals and strategies. For example, in customer projects, we ensure that the chosen goals and strategies are both relevant for the whole organization and not in conflict with other organizational goals and strategies. Strategies towards the solution of the problem may include: (1) understanding the impact of a technology[3], (2) compare alternative technologies, (3) building new technologies, and (4) adopting them in a given context.

The GQM^+Strategies® grid describes the relationships among goals and strategies with rationales. This provides an initial theory for the solution approach. Theory means a set of coherent assumptions (or hypotheses) that describe the expected relation between the selected strategies and the targeted goals. The theory

---

[3] We use the term technology to refer to technique, method, and tool, following Basili et al. [BRC01a].

can be built on both, individual experience of involved experts and empirical evidence.

In addition, we define success criteria, i.e., metrics and target objectives (quantified hypotheses) that allow us to judge whether the selected strategies were successfully implemented.

The motivation of using the GQM+Strategies® approach is to make the relationships between problems, goals, and solution strategies transparent and to provide a starting point for evaluating the solution when it is finally implemented.

Now we are able to describe the overall research question, in particular: What is the problem? Where does it occur? Who has observed it? Why it is important? How to address it (i.e., strategy or solution)? What are the expected results?

Choose Process

After specifying the goals and selecting strategies, the research plan referring to the implementation of the strategies is created. This encompasses among others the order and contexts in which the technologies (i.e., strategies) will be implemented, steps for implementing the technology, associated resources, time schedule for achieving the related goals, and the process for collecting measures. For empirical studies, this means that we select an appropriate design according to the research question (e.g., [ES+08]). Concerning the understanding and comparison of technologies, we use a large range of empirical methods for evaluating these technologies depending on the research question and context of the study. The plans for corresponding empirical studies have to be described.

While developing the research plan the following aspects are considered: The implementation is preferably performed in iterations using formative evaluations to obtain feedback already in early stages. Regarding the adoption of a strategy in an organization, we usually perform first an early evaluation aiming at getting feedback from relevant stakeholders. This evaluation focuses, e.g., on the feasibility of the selected strategy or part of it. For instance, we use focus groups, scenario-based interviews with end users, or small-scaled, well focused experiments. The collected feedback is used for improving the strategy iteratively. A concrete example for this approach is reported in [KJ+09].

In several customer projects, we have observed that industry is reluctant to share necessary data for publications so that we either have to deal with relative numbers or to generally change the measurement approach to a more qualitative approach. In those cases, questions regarding, e.g., acceptance and perceived increase in effectiveness are used [VM+03, VB08] to obtain valuable feedback [NJ+12].

Depending on the nature of the selected strategies, the execution phase may comprise: building models or technologies, understanding the impact of a technology, comparing technologies, introducing them into an organization, collecting and analyzing data regarding goal achievement and further improving the technology.

Models are built to describe reality, respectively a small part of it and either to explain, predict, or prescribe what is, will, or has to be going on. Building technologies include, e.g., defining a process model, developing a new software technology or enhancing an existing one.

It has to be mentioned that the procedure described below accounts for all empirical studies performed in the course of the empirical research process.

According to [CR83] any empirical study has to start with a profound formulation of the research question, which is the basis for any following evaluation step. If the empirical study is part of a larger project, the research question for the study is derived from the overall research question (cf. process step: set goals). A first decision has to be made, whether the research question is of explorative or explanatory nature. In contrast to an explanatory research, exploratory research is conducted in a less fixed environment; many aspects, e.g., variables and their cause-effect relationship are not yet known. The objective is to identify hypotheses which then can be used in explanatory research. For explanatory research, a clear formulation of the research questions towards the evaluation object (e.g., a safety method) and its impact on a certain quality aspect (e.g., consistency of system- and failure model) within a specific context (e.g., avionic domain) from the perspective of the relevant target group (e.g., safety engineer) helps to focus the work and supports a systematic, traceable approach towards the conclusions.

The next step is to further elaborate the underlying theory and investigate related work. Combination of research question, underlying theory, and related work is used to derive hypotheses. There are three types of hypotheses: either they describe a relationship, a difference, or a change in variables. Different types of variables are used to give a model-based description of the environment of interest: independent variables, dependent variables, and confounding variables. Within an empirical study, we assess whether the variation of the independent variable(s) (e.g., safety analysis method) causes an effect in the dependent variable(s) (e.g., consistency). In addition to the independent variable, unknown variables – the confounding variables (e.g., participants experience in safety analysis) – might influence the dependent variable(s). To measure the outcome in the dependent variable(s) requires an operationalization of all variables.

Next, the appropriate empirical research method according to the research question, hypotheses, variables, and given context has to be selected. Typically, empirical methods such as experiments, case studies, and surveys are used. Experiments are used to measure and analyze the effect of systematic variations in the independent variable on the dependent variable. Often an experiment includes an experimental group (treatment applied) and a control group (no treatment applied) to prove the effect of the treatment. For further details on several experimental designs we refer to [SCC02].

If the results of the experiment show that the solution solves the target problem satisfactorily, then we focus on evaluating its potential in real settings. We perform, e.g., case studies in the industry.

According to Yin [Yin03], “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”. Often case studies have small sample sizes and do not allow for controlling confounding variables. Case studies include feasibility studies, observatory single subject design, longitudinal single subject design, and semi-structured feedback sessions. Surveys are used to collect information from a large target group. They provide the data *“to describe, compare or explain knowledge, attitudes and behavior”* [PK01]. Typically used methods are interviews and online-questionnaires.

Furthermore, the target population has to be defined and a (representative) sample has to be drawn. In our example, the target population consists of software and system engineers.

Then, necessary materials, apparatus, and instruments have to be derived and developed; those are the main tools to induce a treatment (e.g., conduct a specific task with a given technology) and to measure the dependent variables (e.g., a questionnaire to measure technology acceptance).

The procedure for running the study has to be fixed. This includes describing precisely what will happen to the participants from the moment they arrive to the moment they leave the experiment site [Har02].

A plan for the summative evaluation of the implemented strategy concludes this phase. Now, we have the plan to execute the research.

Execute and Analyze

The execution aims at implementing the selected strategies according to the specified plan.

For instance, in recent customer projects, we built generic models for quality and costs estimation using a two-step survey approach. We first interviewed a sample of practitioners to provide input on how they characterize quality and costs. The results were used for building initial models. In parallel, we performed a systematic review on the topic. Its results were used for enhancing the initial models. Second, we interviewed the same sample of practitioners for discussing the developed models e.g., [WL+12].

In research projects, we used series of empirical studies such as interviews, focus groups, systematic reviews, and experiments for building information models for supporting decision makers in the selection of software technologies [Jed09], for developing a hybrid method for cost estimation [Tre13], or for analyzing the seamless integration of processes with tools [LRM12].

After implementing the chosen strategies or developing a technology, we usually perform a summative evaluation for assessing to which extend the goals where accomplished.

After a study has been conducted and all data were collected, the results have to be analyzed and interpreted. Feedback sessions with the participants help to better explain the results.

At this point we gained empirical evidence regarding the appropriateness and suitability of the selected strategy in the context at hand. On the basis of this empirical evidence a better decision of whether the technology (i.e., strategy) will be piloted or introduced into the organization or needs to be further improved or changed (new iteration) can be made [Jed10].

For example, in a large German research project, in the context of system safety analysis in the avionics domain, we were asked to compare the currently used state-of-the-practice method for describing a system's safety in terms of a failure model with a model-based method for the same purpose. The focus of the evaluation was on the consistency of the system model with the resulting failure model from the viewpoint of safety engineers. Because of the complexity of the task the study had to be performed with people having respective technical and domain knowledge. In fact, the study was performed with practitioners and replicated with PhD students, who work on projects with a focus on system safety [JH+12]. The results of the studies were well perceived by the industry partner and led to further research questions, which are currently investigated.

Now we have completed the implementation of the strategy and gained empirical evidence regarding its impact in the context of interest.

Package

We distinguish between the packaging of individual studies, the project and synthesis of individual empirical studies.

Individual empirical studies are reported in parallel to the Fraunhofer IESE applied research process. For this purpose, we provide guidelines for different types of empirical studies, e.g., controlled experiments [JPC08], surveys [KP01, CLB03], case studies [RH09], systematic literature reviews [Kit04], and replications [Car10]. Each empirical study has to be placed into the body of knowledge and assessed with regard to the implications of the findings and validity threats.

In the context of the Fraunhofer IESE applied research process, the scope of a project can address a subset or all steps above. The methods and procedures follows during a project as well as its results and lessons learned are summarized in technical reports and project deliverables. These are made available in the Fraunhofer IESE internal repository and are published according to confidentiality agreement signed with the project stakeholders.

Now all empirical studies and their results as well as the project are documented and accessible in the Fraunhofer IESE internal repository. If applicable, empirical studies are synthesized for increasing the body of knowledge of SE.

Once several individual empirical studies have being published regarding a software technology, we emphasize the need to synthesize them for building knowledge in SE systematically. Thus, we adopted existing quantitative and qualitative synthesis methods. For quantitative synthesis, meta-analysis was extended with an approach to incremental aggregation, based on aggregation states, which represent the aggregated knowledge of a set of studies, and which allow to derive the required quantitative key information (e.g., effect existence and magnitude) [Cio12].

Now all empirical studies and their results as well as the project are documented and accessible in the Fraunhofer IESE internal repository. If applicable, empirical studies are integrated into the body of SE knowledge.

3. Discussion and Outlook

We described an empirical process for applied research that (1) aligns our research with problems that are relevant for industry, (2) ensures involvement of relevant sources of information, (3) makes the contribution of the solution towards the problem at hand transparent, and (4) provides us with empirical evidence regarding the appropriateness of the solution. The results are used to support customers making informed decisions [Jed09, Cio12].

Although the request for studies in industry was prevalent from the beginning, the number of publications did not reflect it. Most published studies report on studies with students. The reasons are manifold; among them are: ease of accessing participants, ease of planning, control of material's and procedure's complexity, ownership of the results, ability to publish. The trade-off, as often criticized, comprise limited realism (external validity) and thus acceptability of the results by practitioners [JC+07]. Several authors, aiming at better "marketing" research results, discuss the pros and cons of studies in artificial settings or with student participants [RH09] as well as how to better address the information needs of practitioners [Jed10]. Strategies leading to more studies in industry, are frequently discussed in the empirical SE community, e.g., at the Dagstuhl 2006 Seminar [BR+07] and subsequent ISERN meetings. Those discussions aim at learning from success stories and identifying feasible models of research-industry collaboration. In addition, individual authors report on organizational set-ups for successful technology transfer [Rom00] and models for research-industry collaborations [RA07, BR+07], as well as aspects and lessons learned which have to be taken into account [JP03] such as building trust, confidentiality, and generating a win-win situation.

In our projects with industry, we currently apply the following approaches:

- Individual Studies with practitioners: In close cooperation with an industry partner, we involve practitioners from the business units in a study. This approach is often supported by public funded projects. The major success factor is to motivate the participating organization by convincing them of the potential to gain knowledge. They need to be able to identify a win-situation.
- Research Lab: We invite practitioners to collaborate on a solution to a problem at hand in an environment at the research organization that is close to what they know but can easily be enriched with new technology.
- Consultancy and service provision: We provide services to our customers in terms of supporting them to plan and run empirical studies, analyze the results, and feed them back into the organization.

- Bi-directional exchange: This approach is similar to sabbaticals; in the first phase a practitioner visits a research organization and takes part in the daily work on a topic of his interest. In the second phase the researcher visits the company and works together with the practitioner on the transfer of the solution obtained during the research into practice.

Involving practitioners into studies is also achieved by either paying them [Tic11] or by organizing large coding contests [CAT13].

Outlook

Aiming at supporting efficient and informed decision-making, from a research perspective, we still see the necessity to better understand how and under which circumstances we can synthesize available empirical evidence on particular SE techniques.

Empirical evidence in empirical SE is often created through mixed methods, i.e., the combination of qualitative and quantitative research methods. It can be observed in several research areas, that quantitative and qualitative research are not seen any more as being rivals [LG85] but being complementary. The discussion mainly concerns individual studies, where integration takes place, e.g., by employing triangulation or mixed-method approaches [PMP07]. In empirical SE, to build a comprehensive body of knowledge, synthesizing evidence implies the integration of quantitative and qualitative evidence, which is still an open question in other fields of research [MH94, DWA05].

As for the discussion regarding the applicability of methods for primary studies, the question, whether and why empirical SE is different from other empirical disciplines and consequently requires “own” methods is not answered sufficiently.

In addition, we see the people factor as being particularly important to be further investigated, e.g., what is the impact of experience, knowledge, motivation, and cultural background on the performance in empirical studies. Researcher often claim to have those factors under control by using pre- and post-questionnaires and testing whether they had significant influence on the results. However, we experience issues in some studies we did, especially with practitioners, who seem to underestimate their level of experience for several reasons. We acknowledge that knowledge tests might be too effort prone, but we think that the way of how experience et al. is characterized needs to be revisited and if we go further ahead, a standard for the people factors must be defined.

Acknowledgements

First of all, we thank Dieter Rombach, who provided us with such an inspiring environment as the Fraunhofer IESE. In addition, we would like to acknowledge the contributions of current and former colleagues as well as friends within the ISERN community to the evolution of empirical SE at the Fraunhofer IESE. Due to the number of those, we are not able to list all of them, but Vic Basili, Marcus Ciol-

kowski, Natalia Juristo, and Carolyn Seaman with whom we closely worked together and drove along the empirical SE road for the last 10 years.

References

[APA01] American Psychological Association: *Publication Manual of the American Psychological Association*, (5th ed.). Washington, DC: American Psyc.Assoc, 2001.

[APW02] Aurum, A.; Petersson H.; Wohlin, C.: State-of-the-Art: Software Inspections after 25 Years, *Software Testing Verification and Reliability*, Vol. 12, No. 3, pp. 133-154, 2002.

[BCR01a] Basili, V.R.; Caldiera, G.; Rombach, H.D.: Experience Factory. In: Marciniak J.J. (ed.), *Encyclo. of SE*, Vol.1, John Wiley & Sons, 2001, pp. 511-519.

[BCR01b] Basili, V.R.; Caldiera, G.; Rombach, H.D.: Goal Question Metric Paradigm; in: Marciniak J.J. (ed.), *Encyclo. of SE*, Vol.1, John Wiley & Sons, 2001, pp. 528–532.

[BH10] Basili, V.; Heidrich, J.; Lindvall, M.; Münch, J.; Regardie, M.; Rombach, D.; Seaman, C.; Trendowicz, A.: Linking Software Development and Business Strategy Through Measurement, *IEEE Computer*, vol. 43, no. 4, pp. 57-65, 2010.

[BR+07] Basili, V.R.; Rombach, D.; Schneider, K.; Kitchenham, B.; Pfahl, D.; Selby, R.W. (Eds.): *Empirical Software Engineering Issues: Critical Assessment and Future Directions*, Intern. WS Dagstuhl Castle, Germany, Jan. 2007, Springer LNCS 4336.

[BSH86] Basili,V.; Selby, R.; Hutchens, D.: Experimentation in Software Engineering; *IEEE Trans. on SE*, vol. 12(7): 733-743, July 1986

[Car10] Carver, J.: Towards Reporting Guidelines for Experimental Replications: A Proposal. In Proc of the *1st Intern WS on Replication in ESE Research (RESER) @ ICSE*. May 4, 2010. Cape Town, South Africa.

[CAT13] http://www.catalysts.cc/contest/ (last visted 14.02.2013)

[CD11] Cruzes, D.S.; Dybå, T.: Recommended Steps for Thematic Synthesis in Software Engineering. *ESEM 2011*:275-284

[CD+03] Conradi, R.; Dybå, T.; Sjøberg, D.; Ulsund, T.: Lessons Learned and Recommendations from two Large Norwegian SPI Programmes. In Oquendo (ed.): 9th Europ. *WS. on Software Process Technology* (EWSPT 2003), Helsinki, Finland, 1-2 Sept. 2003, pp. 32-45. LNCS 2786, Springer-Verlag.

[CD+06] Conradi, R.; Dybå, T.; Sjøberg, D.K.; Ulsund, T. (eds): *Software Process Improvement – Results and Experience from the field*. Sringer, Berlin, 2006.

[CHC02] Chalmers I., Hedges L., Cooper H.:A brief history of research synthesis. *Eval Health Prof* March, 25(1):12–37.

[Cio09] Ciolkowski, M.: What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering. ESEM 2009:133-144

[Cio12] Ciolkowski, M.: An Approach for Quantiative Aggregation of Evidence from Controlled Experiments in Software Engineering. Stuttgart: Fraunhofer Verlag, 2012, 231 pages, Kaiserslautern, Univ., Diss., 2011

[CLB03] Ciolkowski, M.; Laitenberger, O.; Biffl, S.: Software reviews: The state of the practice. *IEEE Software*, vol 20 (Nov. Dec. 2003), no 6, 2003, p 46 51.

[CR83] H. T. Chen, P. H. Rossi: Evaluating with sense: The theory-driven approach. *Evaluation Review*, Vol. 7, No. 3, 1983, pp. 283-302.

[Cre03] Creswell, J.W. (2003) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (2nd edition). Thousand Oaks, CA: Sage, 2003.

[CW03] Conradi, R.; Wang, A.I. (eds.): *Empirical Methods and Studies in Software Engineering – Experiences from ESERNET*, Berlin: Springer-Verlag, LNCS 2765, 2003.

[DD08] Dybå, T. and Dingsøyr, T. Empirical Studies of Agile Software Development: A Systematic Review, *Information and Software Technology*, Vol. 50, No. 9-10, Aug. 2008, pp. 833-859.

[DJ11] Dieste, O.; Juristo, N. Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques. *IEEE Trans. on SE*, Vol. 37, no. 2, pp 283-304, 2011.

[DWA05] Dixon-Woods, M.; Agarwal, S.; Jones, D.; Young, B.; Sutton, A.: Synthesising qualitative and quantitative evidence: a review of possible methods*; J Health Serv Res Policy* Jan 2005 10:45—53B

[ES+08] Easterbrook, S. M., Singer, J., Storey, M, and Damian, D. Selecting Empirical Methods for Software Engineering Research. In F. Shull et al. (eds) *Guide to Advanced Empirical Software Engineering*, Springer, 2008.

[Fraunhofer IESE] http://www.iese.fraunhofer.de (last visited 14.2.2013)

[FPG94] Fenton, N.E.; Pfleeger, S.L.; and Glass, R.: Science and Substance: A Challenge to Software Engineers, *IEEE Software*, vol. 11, no. 4, July 1994, pp. 86–95.

[Gla04] Glass, R.L.: Matching Methodology to Problem Domain. In: *Column Practical Programmer in Communications of the ACM* /Vol. 47, No. 5, May 2004, pp. 19-21.

[Har02] Harris, P.: *Designing and Reporting Experiments in Psychology*; Open University Press, Berkshire, UK, second edition 2002.

[ISERN] http://isern.iese.de (last visited 14.2.2013)

[IEEE90] IEEE Computer Society: IEEE: *Standard Glossary of Software Engineering Terminology*. IEEE Standard 610.12-1990.

[JC04] Jedlitschka, A.; Ciolkowski, M.: Towards Evidence in Software Engineering. In: Proc. *Intern. Symp. on Empirical SE 2004* (ISESE2004), Redondo Beach, California, USA, August 2004, 2004, pp. 261-270.

[JC+07] Jedlitschka, A.; Ciolkowski, M.; Denger, C.; Freimut, B.; Schlichting, A.: Relevant Information Sources for Successful Technology Transfer: A Survey Using Inspections as an Example; In: *Proc. Intern. Symp. on Empirical SE and Measurement 2007* (ESEM2007), Madrid, Spain, Sept. 2007, pp. 31-40.

[JCP08] Jedlitschka, A.; Ciolkowski, M.; Pfahl, D.: Reporting Controlled Experiments in Software Engineering; in Shull, F. et al. (eds.): *Guide to Advanced Empirical Software Engineering*; Springer 2008;

[Jed09] Jedlitschka, A.: *An empirical model of software managers information needs for software engineering technology selection*; Stuttgart: Fraunhofer Verlag, 2009, 435 pages. Kaiserslautern, Univ., Diss., 2009

[Jed10] Jedlitschka, A.: Evaluating a model of software managers' information needs: an experiment. In Proc. *ACM-IEEE Intern. Symp. on Empirical SE and Measurement (ESEM '10)*. ACM, Bozen, Italy, Sept. 2010, No. 19, 10 pages.

[JH+05] Jedlitschka, A.; Hamann, D. Göhlert, T.; Schröder, A.: Adapting PROFES for Use in an Agile Process: An Industry Experience Report. In: Bomarius, Frank (Ed.) et al.: in Proc. of *6th Intern. Conf. on Product Focused Software Process Improvement*. Profes'2005; LNCS 3547, 2005, pp. 502-516.

[JH+12] Jung, J.; Höfig, K.; Hiller, M.; Jedlitschka, A.; Domis, D.: *Are PhD-students with domain knowledge appropriate subjects for experiments?* Kaiserslautern, 2012. IESE-Report; 037.12/E

[JM01] Juristo, N. and Moreno, A.: *Basics of Software Engineering Experimentation*; Kluwer Academic Publishers, 2001.

[JM03] Juristo, N. and Moreno, A. (eds.): *Lecture Notes on Empirical Software Engineering*, Ed. River Edge, NJ, USA: World Scientific Publishing, October 2003.

[JMV05] Juristo, N. and Moreno, A.; Vegas, S.: Reviewing 25 Years of Testing Technique Experiments. *Journ. Empirical SE*, vol. 9, No. 1-2, March 2004, pp. 7–44.

[JP03] Jedlitschka, A.; Pfahl, D.: Experience-Based Model-Driven Improvement Management with Combined Data Sources from Industry and Academia. In: Proc. Intern. *Symp. on Empirical SE* ISESE 2003, Roman Castles, Italy, Oct. 2003, pp. 154-161.
[JP04] Jedlitschka, A.; Pfahl, D. (2004): Requirements of a Tool supporting decision making for SE Technology Selection; *In Proc. of Intern. Conf. SEKE*, Banff, Canada, 2004, pp. 513-516.
[JP05] Jedlitschka, A.; Pfahl, D. (2005); Reporting Guidelines for Controlled Experiments in Software Engineering; In Proc. of *Intern. Symp. on empirical SE* 2005 (ISESE2005), Noosa Heads, Australia, Nov 2005, IEEE CS, 2005, pp. 95-104.
[KDJ04] Kitchenham, B.A.; Dybå, T.; Jørgensen, M.; Evidence-based Software Engineering; In Proc. of 26th *Intern. Conf. on Software Engineering* (ICSE'04); May 2004; Edinburgh, Scotland, United Kingdom, 2004, pp. 273-281.
[Ker01] Kerth, N.; *Project Retrospectives: A Handbook for Team Reviews*, Dorset House Publishing, 2001
[Kit04] Kitchenham, B.A.: *Procedures for Performing Systematic Reviews*; Keele University Technical Report TR/SE-0401; ISSN:1353-7776.1 July, 2004.
[KJ+07] Kleinberger, T.; Jedlitschka, A.; Storf, H.; Steinbach-Nordmann, S.; Prueckner, S.: An Approach to and Evaluations of Assisted Living Systems Using Ambient Intelligence for Emergency Monitoring and Prevention; in *Universal Access in HCI. Intelligent and Ubiquitous Interaction Environments*, LNCS 5615, 2009, pp 199-208
[KLL97] Kitchenham, B.; Linkman, S.; Law, D.; "DESMET: a methodology for evaluating software engineering methods and tools," *Computing & Control Engineering Journal* vol.8, no.3, pp.120-126, June 1997
[KN+10] Kläs, M.; Nakao, H.; Elberzhager, E.; Münch, J.: Support Planning and Controlling of Early Quality Assurance by Combining Expert Judgment and Defect Data - A Case Study, *Journ. Empirical SE*, vol. 15, no. 4, pp. 423-454, Springer, 2010
[KP+02] Kitchenham, B.A.; Pfleeger, S.L.; Pickard, L.M.; Jones, P.W.; Hoaglin, D.C.; El Emam, K.; Rosenberg, J.: Preliminary guidelines for empirical research in software engineering; *IEEE Trans. on SE*, Vol. 28, No. 8 , Aug 2002, pp. 721 -734.
[LG85] Lincoln, Y.; Guba, E.G.: *Naturalistic Inquiry*. London, Thousand Oaks, New Delhi: Sage, 1985
[LR97] Linkman, S.; Rombach, D.: Experimentation as a vehicle for software technology transfer-A family of software reading techniques, *Information and Software Technology*, Volume 39, Issue 11, 1997, Pages 777-780,
[LRM12] Lampasona, C.; Rostanin, O.; Maus, H.: Seamless Integration of Order Processing in MS Outlook using SmartOffice: An Empirical Evaluation. In: Proc. of *Intern. Symp. on Emp. SE and Measurement*. New York: ACM Press, 2012, 165-168.
[MH94] Miles, M. B.; Huberman, A. M.: *Qualitative Data Analysis: an Expanded Sourcebook*, Thousand Oaks, Calif., Sage, 1994.
[NJ+11]Nunnenmacher, S. ; Jung, J. ; Chehrazi, G. ; Klaus, A. ; Lampasona, C. ; Webel, C.; Ciolkowski, M.: A Preliminary Survey on Subjective Measurements and Personal Insights into Factors of Perceived Future Project Success; In Proc. of 5th Intern. Symp. on Empirical SE and Measurement. Los Alamitos: IEEE CS, 2011, 396-399.
[Pfl94] Pfleeger, S.L.: Design and Analysis in Software Engineering: Part 1: The Language of Case Studies and Formal Experiments, *ACM SIGSOFT Software Engineering Notes*, Volume 19 Issue 4, October 1994, pp. 16-20.
[PK01] Pfleeger, S.L.; Kitchenham, B.A.: Principles of survey research Part 1: turning lemons into lemonade, *ACM SIGSOFT SE Notes*, Vol 26(6), 2001, pp. 16-18.
[PMP07] Pope, C.; Mays, N.; Popay J.: *Synthesizing qualitative and quantitative health evidence. A guide to methods*. Berkshire, Open University Press. 2007

[RA07] Rombach, D.; Achatz, R.: Research Collaborations between Academia and Industry. In Proc. of *WS on Future of SE 2007*: 29-36

[RH09] Runeson, P. and Höst, M. (2009) Guidelines for conducting and reporting case study research in SE. *Journ. Empirical Softw. Eng.* 14, 2 (April 2009), 131-164.

[RBS92] Rombach, H.D.; Basili, V.R.; Selby, R.W. (Eds.): *Experimental Software Engineering Issues: Critical Assessment and Future Directions*, International Workshop Dagstuhl Castle, Germany, September 1992, Springer Verlag, LNCS 706.

[Ruh03] Ruhe, G.: Software Engineering Decision Support - A New Paradigm for Learning Software Organizations. In Henninger, S.; Maurer, F. (eds): Advances in LSO. *4th Intern. WS, LSO 2002*, Chicago, Il, USA, Aug 6, 2002: Revised Papers LNCS Vol. 2640, Springer 2003, pp. 104-115.

[Rom00] Rombach, D.; Fraunhofer: the German model for applied research and technology transfer, *Proc. of Intern. Conference on Software 2000.* pp.531-537, 2000

[Rom11] Rombach, D.: Empirical Software Engineering Models: Can They Become the Equivalent of Physical Laws in Traditional Engineering? *Int. J. Software and Informatics 5*(3): 525-534 (2011)

[SCC02] W.R. Shadish, T.D. Cook, D.T. Campbell: *Experimental and quasi-experimental design for generalized causal inference*. Houghton-Mifflin, Boston, 2002.

[Sch70] Schulz, W.: *Kausalität und Experiment in den Sozialwissenschaften – Methodologie und Forschungstechnik*. V. Hase & Koehler Verlag Mainz, 1970.

[Sha03] Shaw, M.: Writing Good Software Engineering Research Papers; In Proc. of the *25th Intern. Conf. on SE* (ICSE'03), Portland, Oregon, IEEE CS, 2003, pp. 726-736.

[SD+08] Sjøberg, D.I.K.; Dybå, T.; Anda, B.C.D.; Hannay, J.E.: Building Theories in Software Engineering, In: Shull, F. et al. (eds.); *Guide to Advanced Empirical Software Engineering*; Springer 2008;

[SH+05] Sjøberg, D.; Hannay, J.; Hansen, O.; By Kampenes, V.; Karahasanovic, A.; Liborg, N.-K.; Rekdal, A.: A Survey of Controlled Experiments in Software Engineering. In *Trans. on SE*, Vol. 31, 9, 2005, pp. 733-753.

[Sin99] Singer, J.: (APA) Style Guidelines to Report Experimental Results; In Proc. of *WS on Empirical Studies in Software Maintenance*, Oxford, UK. Sept. 1999. pp. 71-7.

[Tic11] Tichy, W.: Empirical software research: an interview with Dag Sjøberg, University of Oslo, Norway. Ubiquity 2011, June, Article 2 (June 2011), 14 pages.

[Tre13] Trendowicz, A.: *Software Cost Estimation, Benchmarking, and Risk Assessment. The Software Decision-Makers' Guide to Predictable Software Development*. The Fraunhofer IESE Series on Software and Systems Engineering. Springer, Jan 2013

[Tur04] Turner, R.: Why we need empirical information on best practices; *CROSSTALK - The Journal of Defense Software Engineering*, April 2004.

[VB08] Venkatesh, V.; Bala, H. (2008), "Technology Acceptance Model 3 and a Research Agenda on Interventions", *Decision Sciences* 39(2): 273–315

[VM+03] Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D.: User acceptance of information technology. *MIS Quarterly*, (27:3), 2003, pp. 425-478.

[WL+12] Wagner, S.; Lochmann, K.; Heinemann, L.; Kläs, M.; Trendowicz, A.; Plösch, R.; Seidl, A.; Goeb, A.; Streit, J.: The Quamoco Product Quality Modelling and Assessment Approach, Proc. of *34th Intern. Conf. on SE* (ICSE 2012), Zurich, Switzerland, June 2-9th, ACM/IEEE, pp. 1133-1142, 2012.

[WPA03] Wohlin, C.; Petersson, H.; Aurum, A.: Combining Data from Reading Experiments in Software Inspections; In [*JM03*], pp. 85-132.

[WR+00] Wohlin, C.; Runeson, P.; Höst, M.; Ohlsson, M. C.; Regnell, B. and Wesslén, A. *Experimentation Software Engineering*, Kluwer Academic Publishers, 2000.

[Yin03] Yin, R.K.: *Case Study Research*. Design and Methods. 3rd ed. Thousand Oaks : Sage Publications, 2003.