Empir Software Eng (2011) 16:365–395
DOI 10.1007/s10664-010-9146-4

A method for evaluating rigor and industrial relevance of technology evaluations

Martin Ivarsson · Tony Gorschek

Published online: 6 October 2010

Editor: Forrest Shull

Abstract One of the main goals of an applied research field such as software engineering is the transfer and widespread use of research results in industry. To impact industry, researchers developing technologies in academia need to provide tangible evidence of the advantages of using them. This can be done trough step-wise validation, enabling researchers to gradually test and evaluate technologies to finally try them in real settings with real users and applications. The evidence obtained, together with detailed information on how the validation was conducted, offers rich decision support material for industry practitioners seeking to adopt new technologies and researchers looking for an empirical basis on which to build new or refined technologies. This paper presents model for evaluating the rigor and industrial relevance of technology evaluations in software engineering. The model is applied and validated in a comprehensive systematic literature review of evaluations of requirements engineering technologies published in software engineering journals. The aim is to show the applicability of the model and to characterize how evaluations are carried out and reported to evaluate the state-of-research. The review shows that the model can be applied to characterize evaluations in requirements engineering. The findings from applying the model also show that the majority of technology evaluations in requirements engineering lack both industrial relevance and rigor. In addition, the research field does not show any improvements in terms of industrial relevance over time.

Keywords Systematic review · Requirements engineering · Technology evaluation

M. Ivarsson (✉)
Department of Computer Science and Engineering, Chalmers University of Technology, SE-412 96 Göteborg, Sweden
e-mail: martin.ivarsson@chalmers.se

T. Gorschek
School of Engineering, Blekinge Institute of Technology, PO Box 520, SE-372 25 Ronneby, Sweden
e-mail: tony.gorschek@bth.se

1 Introduction

The state of empirical evaluations performed in software engineering research has been criticized over the years (Potts 1993; Glass 1994; Kitchenham et al. 2002). The main arguments center around a lack of empirical evaluations (Tichy et al. 1995; Glass et al. 2002), poor study execution (Kitchenham et al. 2002; Dybå et al. 2006) and a lack of realism when performed (Sjøberg et al. 2002). This may hamper progress in the field, as practitioners looking to adopt new technologies developed in academia are offered scarce decision support (Ivarsson and Gorschek 2009). In an applied research field such as software engineering, the transfer and widespread use of research results in industry ultimately determine the relevance and success (Sjøberg et al. 2002; Gorschek et al. 2006). Researchers should thus not only develop new technologies and advocate potential benefits but also base the information on the extent to which the technology has been validated (tested). A lack of realistic evaluations also limits the potential for practice to influence research, as evaluations carried out in industry provide feedback on what works in practice. In this context, a technology can be seen as any method, technique, procedure or model (or combination thereof) used in software development and maintenance (Pfleeger 1999; Pfleeger and Menezes 2000; Ivarsson and Gorschek 2009).

Practitioners looking for new technologies must be able to evaluate the potential benefits and risks prior to adopting them. For the evidence to be convincing to industry, evaluations must be performed in a realistic setting, using realistic subjects and tasks, not using only toy examples, as practitioners value evaluations in settings comparable to their own environments (Zelkowitz et al. 1998). The realism of evaluations thus determines the potential relevance of research results for industry. The trustworthiness of results must be considered in addition to relevance. The rigor of an evaluation and the way evaluations are presented influence the amount of trust placed in the evidence. Usefulness is limited without a description of how the results are obtained or in what context they are valid.

This paper presents a model for evaluating rigor and the industrial relevance of technology evaluations in software engineering. Traditionally, the relevance or impact of research results is associated with, for example, the number of citations (Wohlin 2009a). However, the impact on research in the form of the number of publications or citations does not necessarily reflect success in the field. The model presented in this paper aims to balance this view of the relevance of research by evaluating the potential for impacting industry. The potential for impacting industry is evaluated by considering how the results are obtained and the realism of the setting (Sjøberg et al. 2002; 2007). Practitioners value results obtained in settings similar to their own with research methods relevant in their environment (Zelkowitz et al. 1998). In addition to relevance, the trustworthiness of the results presented is also considered. Rigor is evaluated indirectly by gauging to what extent and detail the context of a technology evaluation is presented by researchers. Failing to report aspects related to rigor makes it difficult to understand and use the results and limits the possibility for other researchers to replicate or reproduce the study (Wohlin et al. 2000).

The model can be used in several ways.

- It can be used to characterize and evaluate the level of rigor and relevance in a research field and how it has progressed over time.
- Practitioners looking for new technologies can use the results of the model to gauge the relevance of available evidence in a field and pinpoint studies potentially relevant for them.
- Researchers looking for an empirical basis on which to build new or refined technologies or studies to synthesize or replicate can use the results to find well described studies.

- The results of applying the model also highlight gaps or imbalances in research and can motivate the directions of future research.
- The model emphasizes aspects that are important in planning and reporting evaluations in software engineering and can thus be used preemptively when planning and reporting studies. For research to influence practice, the studies that are performed must produce results that are relevant to industry.

To make a pilot examination and provide an initial validation of the model, it is used as a basis for a systematic literature review (Kitchenham and Charters 2007) of technology evaluations in requirements engineering. Requirements engineering research is suggested to have limited impact on industry practice (Juristo et al. 2002; Kaindl et al. 2002; Morris et al. 1998). Thus, investigating the level of relevance and rigor of requirements engineering research, and how it has changed over time can provide insight into one factor potentially influencing this shortcoming.

The review has two purposes and contributions. First, it is used to test and refine the model itself that is aimed at classification of rigor and relevance of technology evaluations. Second, the results that the model produces indicate the level of rigor and relevance of technology evaluations in requirements engineering, and how these have changed over time.

The paper is structured as follows. Section 2 introduces relevance and rigor in software engineering and presents related work. The model for evaluating rigor and relevance of technology evaluations is presented in Section 3. The design of the systematic review and the research questions addressed are given in Section 4, and threats to validity are discussed in Section 5. The results of the review are presented in Section 6, and the validation of the model is discussed in Section 7. Future work is presented in Section 8. Section 9 gives conclusions.

2 Background and Related Work

This section introduces relevance and rigor in software engineering in Sections 2.1 and 2.2, respectively, and explains the use of these terms in this paper. Section 2.3 presents related work on rigor and relevance.

A wide gap has been identified between what is practiced in industry and what is proposed by research (Hsia et al. 1993; Potts 1993; Neill and Laplante 2003). At the heart of this gap is the transfer of research results into use in industry. This paper considers one potential factor governing this transfer by focusing on how technologies are evaluated in research. It also considers how these evaluations can be used by other researchers and, equally important, by industry professionals to gauge the relevance and validity of the evaluations. In short, can researchers get valid results they can build on and/or do industry professionals get enough decision support to dare to try out research results?

Rigor refers to both how an evaluation is performed and how it is reported. If the study is not adequately described, the rigor of the evaluation cannot be evaluated by reviewers and other researchers. A background to and details regarding the rigor of software engineering research is given in Section 2.2. Relevance on the other hand refers to the potential impact the research has on both academia and industry. Academic relevance, or impact, is shown through the ability to publish papers and through citations by other researchers. Industrial relevance concerns the evaluation's value for practitioners seeking to adopt technologies. To aid the transfer of technologies from academia to industry, research

needs to provide evidence of the benefit of adopting research results (technologies). Here, evaluations can provide an incentive for practitioners to try new technologies. A background to relevance in software engineering is given in Section 2.1. Work related to rigor and relevance in software engineering is presented in Section 2.3.

2.1 Relevance

Several different measures have been used to evaluate academic relevance or impact in research, including the impact of individual researchers based on the number of publications (Wong et al. 2009) and different indexes (h-index (Hirsch 2005), w-index (Wohlin 2009b) etc.) related to the number and distribution of citations of researchers or individual articles. This indicates one aspect of the impact on research. However, quantity is only one aspect of impact and critics have argued that it should not be emphasized over quality or the value of research (Parnas 2007). Assessing research quality or the value of research introduces subjective bias in both what to assess, e.g. what constitutes value and quality, and reviewers' competence to assess it (Meyer et al. 2009).

The impact on industry or the industrial relevance of research is even harder to gauge than academic relevance. One decisive factor for the relevance of research results is the actual topic studied (Sjøberg et al. 2007). Studying relevant topics increases the likelihood of the results being relevant to industry. However, the time perspective needs to be considered. Redwine and Riddle found that it takes in the order of 15–20 years to mature a technology to the stage that it can be disseminated to the technical community (Redwine and Riddle 1985). This makes it infeasible to use the actual uptake of research results in industry as the only evaluation tool. It also makes it difficult to distinguish relevant topics from irrelevant ones.

The industrial relevance of software engineering research has been questioned over the years (Potts 1993; Glass 1994). The ultimate goal of an applied research field such as software engineering is the transfer and use of results in industry. The actual impact on industry is not evaluated in this paper but rather the potential for impact. Zelkowitz et al. (1998) found that practitioners value research methods that are most relevant to their own environment, i.e. research methods that concern the performance of technology in live situations. Given the present state-of-research, research methods such as case studies and lessons learned have a greater potential to provide results compelling for industry since experiments are often conducted in unrealistic environments (Sjøberg et al. 2005). This means that evaluations carried out using these research methods should have a greater potential to provide compelling evidence to practitioners. However, the research method is just one aspect that influences relevance. It has also been argued that an increase in the realism of evaluation is central to producing results that support a transfer to industry (Sjøberg et al. 2002, 2007). Evaluations being representative of the target to which to generalize to support the transfer of results (Sjøberg et al. 2007). Realism can be improved by making evaluations that more closely resemble industrial settings. Aspects that must be considered in order to increase realism are the scale, context and subjects used in the evaluation. Scale refers to the time scale and size of the application/use/test used in the evaluation. Using small examples over a short period of time might not show how the technology under evaluation would scale to an industrial setting (Sjøberg et al. 2007). Context refers to the environment in which the evaluation is carried out. In order for practitioners to evaluate the suitability of a particular technology, the context in which it has been used/tested must be described (Pfleeger and Menezes 2000; Ivarsson and Gorschek 2009). This enables an understanding of whether the technology would be suitable for use in another context. If evaluations are made in an

artificial setting, e.g. the classroom, it is difficult to value the results (Sjøberg et al. 2007) and the environments in which they are valid. An artificial setting is also likely to influence the motivation and goals of the subjects involved in the evaluation, thus introducing threats to construct validity (Wohlin et al. 2000). Subjects used in the evaluation have a potential impact on the results obtained as they might not behave in the way professionals would and thus not give results that are valid in industry (Arisholm and Sjoberg 2004; Arisholm et al. 2007).

2.2 Rigor

Rigor in research often refers to the precision or exactness of the research method used; e.g. a controlled experiment often enables greater control over variables than a case study (Wohlin et al. 2000). This is one way to view rigor, the precision of the research approach utilized. Rigor can also mean the correct use of any method for its intended purpose (Benbasat and Zmud 1999; Wieringa and Heerkens 2006), implying that there is a context or application in which certain methods are appropriate or applicable. Several studies have evaluated the rigor used in software engineering research, of which a selection is presented in Section 2.3. The aim of this paper is not to argue for a specific research methodology but rather to emphasize that the methodology used should be carried out in accordance with corresponding best practices. Rigor consists of two components, the rigor of the study and the way it is presented. Reference literature for research methods common in software engineering is given in Table 1. These research methods all have different traits and procedures for collecting and analyzing data, and thus different criteria for what constitutes the rigor of a study. Experimentation and case study research in software engineering have been extensively discussed while surveys and action research in software engineering have received less attention. However, action research has been given much attention in the related research field of information systems research.

In addition to the references in Table 1, guidelines for how to carry out and report empirical work in software engineering can also be found in (Kitchenham et al. 2002). However, rigor not only relates to the way in which a study is carried out. If a study is not presented adequately, a reviewer or other reader cannot determine whether the study has been carried out in a rigorous way, and the rigor of the study thus becomes irrelevant as it is not presented. Failing to report the study in an adequate way also limits the opportunity for other researchers to understand and replicate or reproduce the study (Wohlin et al. 2000). This means that presenting studies in an adequate way is a common denominator for all research methods. A compilation of aspects that should be included in presentations of evaluative research is given in Table 2.

Reference literature shows a high degree of coherence as to what aspects that are important to report in evaluative research. Guidelines for action research in (Lau 1999) do

Table 1 Reference literature for software engineering

Research method	Reference literature
Experimentation	(Basili et al. 1986; Pfleeger 1994–1995; Wohlin et al. 2000)
Survey	(Kitchenham and Pfleeger 2001–2003)
Case study	(Kitchenham 1996–1998; Glass 1997; Yin 2008)
Action research	(Lau 1999)

Table 2 Study presentation aspects

Aspect	Description	References
Related work, research goal, problem statement, hypothesis	Defines the problem/goal of the evaluation and how it relates to previous research	(Lau 1999; Perry et al. 2000; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)
Context	Information about the context in which the evaluation is performed including e.g. the experience of the staff, development process used etc.	(Lau 1999; Perry et al. 2000; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)
Study design	Describes the products, resources and process used in the evaluation e.g. population, sampling etc.	(Lau 1999; Perry et al. 2000; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)
Validity, limitations	Discusses any limitations or threats to the validity of the evaluation including measures taken to limit these	(Perry et al. 2000; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)
Study execution	Describes the execution of the evaluation including data collection, preparations etc. Any sidesteps from the design should also be described. The execution should describe aspects that ease replication.	(Lau 1999; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)
Analysis	A presentation of the analysis where the analysis model and tools are described, as are assumptions for these. In addition, information about significance levels and the applicability of tests should be discussed.	(Lau 1999; Perry et al. 2000; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)
Presentation and interpretation of results	An interpretation of the analysis in relation to the hypothesis/problems addressed by the evaluation	(Lau 1999; Perry et al. 2000; Wohlin et al. 2000; Kitchenham et al. 2002; Yin 2008)

not explicitly state the validity or limitations of a study, but several elements in the guidelines relate to similar concepts such as the credibility of data, degree of openness etc. (Lau 1999). Thus there seems to be a consensus as to what aspects it is important to address when research studies are presented.

2.3 Related Work

Several studies have investigated how software engineering research is carried out with respect to relevance and rigor. A selection is shown in Table 3.

To some extent, previous studies give mixed messages about the state of rigor and relevance in software engineering. In terms of relevance, the research method used has been investigated in several literature reviews. Glass et al. (2002) point out that applying the technology to an example is the dominant research method. However, both Zelkowitz (2009) and Zannier et al. (2006) show that the number of validations, in contrast to pure theoretical or advocating research, has increased over time, indicating an improvement of the state of evaluations. The research method used also seems to depend on the publication venues included in the review. For example, in the Empirical Software Engineering (EMSE) journal, it is reported that experiments and case studies are the most often used research methods (Höfer and Tichy 2007) while results in the Requirements Engineering journal (REj) show that illustrating technologies in an example is the most commonly used research method (Ivarsson and Gorschek 2009). The results are again different when the

Table 3 Literature studies related to relevance and rigor in software engineering

	Scope	Relevance	Rigor
(Sjøberg et al. 2005; Dybå et al. 2006; Hannay et al. 2007; Kampenes et al. 2007)	Controlled experiments from selected joulrnals and conferences	- The majority of subjects used in experiments are students	- Internal and external validity is often reported
		- The majority of applications used in the experiments are constructed for the purpose for the experiment or constitute student projects	- Relatively low and arbitrary reporting of context variables
		The duration of tasks in the experiments are short	- Reporting of validity is often vague and unsystematic
			- Most experiments do not relate to theory
			The level of statistical power falls substantially below accepted norms
(Höfer and Tichy 2007)	Journal of Empirical Software Engineering	- Professionals are used frequently in evaluations	
		- Long-term studies are missing	
		Narrow focus of topics empirically investigated	
(Ivarsson and Gorschek 2009)	Requirements Engineering Journal	- Evaluations of requirements technologies offer scarce support for practitioners looking to adopt technologies	- Experiments using human subjects are on average well described, i.e. present context, study design and the validity of the study
		- Most evaluations are performed by the researcher him/herself	Except for when experiments are performed, few studies describe the context, study design and validity of the evaluations
		- The most used evaluation method is to apply a technology to an example	- The most common type of sampling used in evaluations is investigator sampling
		- Toy examples are used in half of the evaluations	- No significant illegal use of analysis on scales of measurement
		A majority of the evaluations are carried out in academia	- Great misuse of the term “case study”
(Zannier et al. 2006)	ICSE	- An increase in the number of empirical evaluations over time	- Self-confirmatory studies
		- Experience reports and pseudo controlled studies	- The soundness of studies does not

Table 3 (continued)

	Scope	Relevance	Rigor
		are the most common study types	improve over time
		- Use of examples commonly employed as validations	- The type of study performed is only defined in half of the cases
		Half of the evaluations use professionals as subjects	There is a lack of replicated studies
(Zelkowitz and Wallace 1997; Zelkowitz 2009)	ICSE, TSE, IEEE SW	- The amount of validation is increasing	
(Glass et al. 2002)	Selected software engineering journals	- Research approach and method used in software engineering are quite narrow	
		Illustration of the use in examples is the most prevalent evaluation method	

International Conference on Software Engineering (ICSE) is considered, where experience reports and pseudo controlled evaluations are in the majority (Zannier et al. 2006).

Regarding the subjects used in evaluations, and thus to some extent the realism of evaluations, it seems that the results depend on the research method used and the publication venue. For example, Höfer and Tichy (2007) report that the majority of case studies uses professionals as subjects while experiments often utilize students. Sjøberg et al. (2005) also found that subjects are often used in experiments in software engineering. Looking at the publication venue, most of the evaluations in REj are carried out by the researcher him/herself (Ivarsson and Gorschek 2009) while half of the evaluations presented in ICSE use professionals as their subjects.

Scale, which is also an aspect of relevance, has been found to be unrealistic. For example, considering experiments in software engineering, the applications used are often constructed for the purpose of the experiment and the duration of tasks performed in the experiments is short (Sjøberg et al. 2005). In addition, more than half of the evaluations reported in REj is performed in toy or down-scaled examples (Ivarsson and Gorschek 2009).

The rigor of evaluations in software engineering research has also been investigated. First, several studies have found that presentations of research show room for improvement (Kampenes, Dybå et al. 2007; Ivarsson and Gorschek 2009). This includes failing to report both certain aspects, such as validity, study design etc., and the structure and contents of the presentation (Sjøberg et al. 2005; Zannier et al. 2006; Ivarsson and Gorschek 2009). For example, Zannier et al. (2006) found that it is common for researchers not to report the type of study that is carried out, and the REj review showed that aspects related to study design, validity and context are seldom reported except for in experiments.

Detailed aspects concerning the rigor of experiments carried out in software engineering have also been investigated. The actual rigor of experiments reported in software engineering seem to be lacking, as statistical power falls below accepted norms in related fields (Dybå et al. 2006). Only 6% of the studies investigated in (Dybå et al. 2006) had a power of 0.80 or more. Achieving only a weak statistical power makes it impossible for researchers to draw conclusions from experiments.

3 A Model for Evaluating Rigor and Industrial Relevance in Technology Evaluations

A model that captures rigor and the relevance of technology evaluations in software engineering research was developed to evaluate these dimensions. This section gives an overview of the model and details its constituents. The aim of the model is to enable a classification of individual evaluations in order to characterize research carried out in a field. Thus the model needs to be applicable to many types of evaluations, i.e. research methods. The way in which research is classified is described in Section 3.1. The resulting classifications are quantified and visualized to be able to characterize and understand technology evaluations in a given field of research. The quantification into variables for rigor and relevance is described in Section 3.2 and the visualizations employed are given in Section 3.3. Limitations of the model are given in Section 3.4.

3.1 Classifying Research

To analyze the state of technology evaluations, studies are classified from the perspectives of relevance and rigor. The purpose is not to achieve an exact classification of each

individual study but rather to give an approximate overview of the state and progress of research in order to identify patterns. Thus, a detailed in-depth analysis of studies does not necessarily add any benefit over a more simplified one. For example, several aspects might be considered when classifying the extent to which the context is reported, i.e., product, process, practices, tools, techniques, people, organization and market (Petersen and Wohlin 2009). Classifying the reporting of context can then be done by considering the number of aspects reported in each paper. Using this detailed classification would lead to a precision in the results, i.e. a classification that can pinpoint what aspects are included or missing from reports. A high degree of precision in classification is not necessarily sought, however, as the goal of the model is to provide an overview of the state-of-research. A detailed classification would also lead to few (if any) papers being classified at the highest level (containing all aspects) (Petersen and Wohlin 2009). To characterize state-of-research, the classification should relate to how research is currently performed and reported, considering only aspects that are often reported. If no papers include a particular aspect, the ability of that aspect to describe the state-of-research is limited. The evolution of the model to improve the ability to capture the state-of-research is described in Section 7.1. Instead of classifying details, a simplified way of classifying relevance and rigor using scoring rubrics is used in the model. Rubrics are one way to formalize the evaluation into different criteria and their levels (Moskal 2000). They have been previously proposed and successfully used in Software Engineering for the evaluation of Master Theses (Feldt et al. 2009). The scoring rubrics used to evaluate rigor and relevance are presented in Sections 3.1.1 and 3.1.2 respectively.

3.1.1 Rigor

It is not the actual rigor of studies, e.g. use of a correct analysis method on the scales of measurement, appropriate sampling type etc., that is considered in this model but rather the extent to which aspects related to rigor are presented. Failing to report aspects related to rigor makes evaluation by reviewers or readers difficult. The focus is on the way in which aspects are reported enabling the classification of diverse study types; e.g. characteristics of rigor for an action research study are not the same as an experiment. Table 4 shows the aspects considered for evaluating rigor of studies and the scoring rubrics that was developed to guide the evaluation. Three aspects are considered in scoring rigor; the extent to which context, study design and validity are described (see Table 4). All these aspects are scored with the same three score levels; “weak”, “medium” and “strong” description. Applying the model uncovered that many papers mentions aspects related to rigor but not describe these fully. Using three levels for judging aspects related to rigor enables identifying these cases. The scoring of each aspect is described in detail in Table 4.

3.1.2 Relevance

The industrial relevance of an evaluation consists of two parts in the model. First, the realism of the environment in which the results are obtained influence the relevance of the evaluation. Three aspects of evaluations are considered in evaluating the realism of evaluations: subjects, scale and context. The scoring rubrics used to assess these are given in Table 5. The expertise or skill of subjects used in the evaluation is also likely to influence the results (Wohlin et al. 2000; Arisholm and Sjoberg 2004; Arisholm et al. 2007; Sjøberg et al. 2007). As the reporting of background and the skill level of subjects used in evaluations varies substantially (Sjøberg et al. 2005); the only aspect included in the model differentiates whether subjects are practitioners, students or researchers.

Table 4 Scoring rubric for evaluating rigor

Aspect	Strong description (1)	Medium description (0.5)	Weak description (0)
Context described	The context is described to the degree where a reader can understand and compare it to another context. This involves description of development mode, e.g., contract driven, market driven etc., development speed, e.g., short time-to-market, company maturity, e.g., start-up, market leader etc.	The context in which the study is performed is mentioned or presented in brief but not described to the degree to which a reader can understand and compare it to another context.	There appears to be no description of the context in which the evaluation is performed.
Study design described	The study design is described to the degree where a reader can understand, e.g., the variables measured, the control used, the treatments, the selection/sampling used etc.	The study design is briefly described, e.g. "ten students did step 1, step 2 and step 3"	There appears to be no description of the design of the presented evaluation.
Validity discussed	The validity of the evaluation is discussed in detail where threats are described and measures to limit them are detailed. This also includes presenting different types of threats to validity, e.g., conclusion, internal, external and construct.	The validity of the study is mentioned but not described in detail.	There appears to be no description of any threats to validity of the evaluation.

To address aspects such as the scalability and usefulness of technologies under evaluation, applications used in the evaluation must have a realistic scale (Sjøberg et al. 2002, 2007). The scale aspect in the model refers to the type of application used in the evaluation and ranges from toy examples to industrial scale applications.

Finally, the context in which results are obtained determines the type of study. For example, a case study in industry is likely an on-line evaluation in a realistic setting, while a similar evaluation performed in academia is probably not.

Second, the research method used to produce the results influence the relevance of the evaluation. A diverse set of research methods is included in the model to cover a wide range of activities from application (test/illustration) of a technology to experiments and any sort of empirical evaluation. The simplest form of evaluation is chosen to be a technology applied to an example. The reason for including application examples as a type of evaluation in the classification is to provide a point of comparison in relation to other research methods. Excluding application examples would remove the ability to appreciate the proportions between different types of evaluations used in software engineering. In addition, researchers in software engineering often use example applications to "validate" or evaluate technologies (Zannier et al. 2006). The scoring discerns research methods that

contribute to relevance from ones that do not. This valuation is likely to evolve over time to reflect how research methods are valued by practitioners, for example, practitioners value studies carried out in the form of case studies or lessons learned (Zelkowitz et al. 1998). This can depend on the state of validation in software engineering, as experiments are often carried out using students experimenting with toy examples. If experiments more closely resembled the environment to which the results are generalized (often industry), this view might change, i.e. practitioners' valuation of experiments could improve if the realism of experiments improved. Thus, scoring should change over time to correspond to how research is carried out. The scoring rubric used to classify research method is detailed in Table 5.

3.2 Quantification and Measurement Conversion

To analyze the resulting classifications in order to present an abstract view of the state of the technology evaluation, the classification is converted into numerical values. This conversion (from a nominal to an ordinal scale) adds information by placing a value on different classifications. Conversion from a weaker to a stronger measurement scale is

Table 5 Scoring rubric for evaluating relevance

Aspect	Contribute to relevance (1)	Do not contribute to relevance (0)
Subjects	The subjects used in the evaluation are representative of the intended users of the technology, i.e., industry professionals.	The subjects used in the evaluation are not representative of the envisioned users of the technology (practitioners). Subjects included on this level is given below: • Students • Researchers • Subject not mentioned
Context	The evaluation is performed in a setting representative of the intended usage setting, i.e., industrial setting.	The evaluation is performed in a laboratory situation or other setting not representative of a real usage situation.
Scale	The scale of the applications used in the evaluation is of realistic size, i.e., the applications are of industrial scale.	The evaluation is performed using applications of unrealistic size. Applications considered on this level is: • Down-scaled industrial • Toy example
Research method	The research method mentioned to be used in the evaluation is one that facilitates investigating real situations and that is relevant for practitioners. Research methods that are classified as contributing to relevance are listed below: • Action research • Lessons learned • Case study • Field study • Interview Descriptive/exploratory survey	The research method mentioned to be used in the evaluation does not lend itself to investigate real situations. Research methods classified as not contributing to relevance are listed below: • Conceptual analysis • Conceptual analysis/mathematical • Laboratory experiment (human subject) • Laboratory experiment (software) • Other • N/A

usually not permissible. However, in this case, the goal is to add information to discern studies that have low rigor/relevance from studies that have higher levels. For example, using professionals in studies is a more realistic solution than using students (Sjøberg et al. 2002, 2007), and thus a higher value is assigned to studies that use professionals.

The value of the constituents is summed up to form variables for rigor and relevance that can be analyzed, as can be seen in Tables 6 and 7. Generally, an addition of measures on the ordinal scale makes little sense. However, additions can be used as long as the resulting scale can be interpreted and used. The resulting variables should be interpreted as how many aspects contribute to industrial relevance for relevance (see Table 7) and how many aspects are described for rigor (see Table 6). For example, a study that is assigned a relevance value of "2" has two aspects that are classified as contributing to industrial relevance. This does not imply however that a study that has twice the value for relevance is twice as likely to influence industry. It only provides an approximation of how many aspects contributing towards relevance a study has.

The ability of this quantification to reflect the actual relevance might be argued. For example, this would rate the relevance of an experiment using students as subjects experimenting with a toy example in academia lower than an application example on an industrial scale carried out by a researcher in academia (as can be seen in Table 7). In this case, the experiment that uses students might produce more relevant results as it does not employ an *ad hoc* research method and is, for example, more likely to control for researcher vested interest. The variable for relevance only illustrates how many aspects of the study are realistic and provides an approximation of the relevance of studies carried out in software engineering.

The average of these variables is also used in the analysis. The average of relevance should be interpreted as how many aspects contributing to relevance are included on average in the studies. The average for rigor should be interpreted as the average number of aspects that are described in the studies. For example, if the average for rigor is "1", this means that, on average, one aspect is fully described or two aspects are mentioned in the studies.

3.3 Visualization

Three different charts are used to visualize the resulting variables, where a bubble chart depicts the combination of rigor and relevance to characterize the type of research carried out in a research field. Bubble charts can be very effective in giving visual indications of how much research is carried out and categorizing research type (Petersen et al. 2008;

Table 6 Quantification of rigor

Rigor, Rigor = C + S + V	
Context described (C)	
Study design described (S)	
Validity discussed (V)	
Each aspect is scored according to the following scheme	
Weak presentation	0
Medium presentation	0,5
Strong presentation	1

Table 7 Quantification of relevance

Relevance, Relevance = C + RM + U + S
Context (C)
Research method (RM)
User/Subject (U)
Scale (S)
Aspects are scored as 1 if contributing to relevance, 0 otherwise.

Šmite et al. 2010). Bubble charts are used to indicate the amount of evaluations that end up with a specific combination of rigor and relevance in Section 6.1. A line chart showing how the average rigor and relevance have changed over time is used to analyze the evolution of research in Section 6.2. Finally, to investigate the influence of the publication venue on the rigor and relevance of the evaluations presented, the average rigor and relevance for the publication venue is considered in Section 6.3.

3.4 Limitations

The goal of the model is to describe state-of-research and to identify studies with high levels of rigor and industrial relevance. Classifying research presentations is inherently an approximation as the classification is of the presentation as opposed to the actual research (Glass et al. 2002) This introduces a number of limitations with respect to using the model:

- The results from the model do not consider the actual results of the evaluated studies. Studies ending up as having high levels of rigor or relevance might describe negative results. The results from using the model only present the level of rigor and relevance of studies included.
- The model does not consider the alignment of the technology being evaluated with the context in which it is being evaluated. For example, a technology developed to be used in large teams can be evaluated in small teams. It is up to the user of the results to value the contribution of the individual papers. The model only provides a way of discerning studies with different levels of rigor and relevance.
- Evaluations of technologies on industry data by the researcher him/herself is classified as having "researcher" as subject. In some cases this mean that the study get a low relevance score relative to its actual value for practitioners. This is a trade-off made in the model to avoid introducing subjective bias in the scoring and to keep the focus on identifying evaluations that take practitioners valuation of usability and usefulness into account.
- The classification of rigor only considers if aspects are presented in the paper. It does not consider if the study design is appropriate to address the research questions or if all threats to validity are disclosed and handled, e.g. a classification of high for the validity aspect only mean that all relevant types of threats are presented.

4 Model Validation—Systematic Review Design

To validate the model presented in Section 3, it is applied in a full systematic literature review of technology evaluations in requirements engineering. This section gives a detailed

account of the design of the systematic literature review. The review has two purposes. First, it illustrates how the model can be used to evaluate the rigor and relevance of technology evaluations in the field of requirements engineering (see e.g. (Afzal et al. 2008)). Second, the results of the review give an overview of the state of rigor and relevance of technology evaluation in requirements engineering (see e.g. (Afzal et al. 2009)). The review follows the guidelines proposed by Kitchenham and Charters (2007). The main deviation from the proposed procedures is the lack of study quality assessment (Kitchenham and Charters 2007), as all studies presenting a technology evaluation are included in the review. Thus, the quality assessment is part of the inclusion criteria and scoping, i.e., the assessment of the quality of the studies included is part of the data extraction procedure.

4.1 Research Questions

The research questions addressed in this review are derived from the model presented in Section 3 and are:

RQ1. What is the state-of-practice in relation to rigor and relevance of technology evaluations in requirements engineering?
RQ2. Have rigor and relevance changed over time?
RQ3. Does publication venue influence rigor and relevance?

The first research question concerns characterizing the type of evidence produced in evaluations in requirements engineering research (relevance) and how it is presented (rigor). Relevance is characterized by the realism of evaluations. The research method used, as well as subjects, context and scale of evaluations, determine the realism of the evaluations. The presentation of evaluations as regards study design, validity and context determines the level of rigor.The second research question seeks to investigate whether there have been any improvements with respect to relevance and rigor over time. This could indicate that technologies and research in the area are maturing to the point where industry trials and applications are more dominant. The last research question investigates the influence of publication venue on relevance and rigor.

4.2 Inclusion Criteria

This review aims to investigate the rigor and relevance of technology evaluations in requirements engineering. The principal criterion for including a paper is thus that it should present an evaluation of a requirements engineering technology. The set of inclusion criteria is detailed below:

- The paper should be in the scope of requirements engineering. A requirement describes a condition or capability needed by the user to solve a problem or achieve an objective (IEEE 1990). There is vagueness between requirements and design, making it hard to define a clear inclusion criterion. However, requirements primarily concerns what the systems should do, not how to do it (Ross and Schoman 1977). Papers that focus on what the system should do are thus included, even if they contain elements of design.
- The paper should present a technology. In this context a technology can be anything from methods, techniques and procedures to models and tools (or combinations thereof) (Pfleeger 1999; Pfleeger and Menezes 2000; Ivarsson and Gorschek 2009).

- The paper should present an evaluation. Evaluation is here defined to cover a wide range of activities from application (test/illustration) of a technology in a toy example invented by the researchers themselves to experiments and any sort of empirical evaluation.

These broad inclusion criteria make for an "include heavy" selection, thus avoiding dismissing papers that have some sort of evaluation of a requirements technology.

4.3 Identification of Articles

A database search was made to identify papers that should be included in the review. The search term was devised to find requirements engineering technologies that have been evaluated and has been used in a previous review (Ivarsson and Gorschek 2009) (see Table 8).

The search term was applied to all journals classified as software engineering by ISI (listed in Table 9). The reason for including papers exclusively from journals is that journals are the premier publication venue in software engineering. Thus they should include the most mature research in the field implying higher degree of empirical evidence than research presented in conferences or workshops. In addition, few journals have hard size restrictions on papers implying that aspects related to rigor is more likely to be included. The database search was done using Inspec, as all included journals are indexed, and applied only to the title, abstract and keywords, as a full text search would cover too many irrelevant results (Dybå et al. 2007). The search was performed on October 16, 2008, and identified 3,593 papers. To identify and extract papers relevant for the review, the title and abstract were read and compared to the inclusion criteria. This identified 455 papers that were included for data extraction. On the basis of reading the full paper, 349 papers were included and classified in the review. The number of papers in each journal that was included is given in Table 9.

4.4 Data Collection

The papers included in the review were analyzed and classified according to the aspects and scoring rubrics in the model presented in Section 3. All aspects are classified according to what is mentioned in the articles that were reviewed. If the aspects are not explicitly mentioned in the paper, it is mapped according to the reviewer's understanding. An article can describe several studies in which case the primary study, as portrayed by the paper, is classified. If that cannot be determined, the study most advantageous with respect to subsequent analysis is chosen. The data extracted were stored in a MySQL database for later analysis.

5 Threats to Validity

The main threats to validity in this study are publication and selection bias, and data extraction, each detailed below.

Table 8 Search term

Population	requirement*, specification
Intervention	empiric* OR experience* OR "lesson learned" OR "lesson learnt" OR "lessons learned" OR "lessons learnt" OR evaluat* OR validation* OR experiment* OR stud* OR case* OR example* OR survey OR analys* OR investig* OR demonstrate*

Table 9 Journals included in the review

Journal	Abbreviation	Papers included
Communications of the ACM	COMACM	4
Computer	COMPUTER	4
Computer Journal	COMPUTERJOURNAL	9
Empirical Software Engineering	EMSE	15
IBM Systems Journal	IBMJOURNAL	1
IEEE Software	IEEESOFTWARE	40
IEE Proceedings-Software	IEEPROC	12
International Journal of Software Engineering and Knowledge Engineering	IJSEKE	19
Information and Software Technology	IST	37
Journal of Research and Practice on Information Technology	JRPIT	5
Journal of Systems and Software	JSS	32
Requirements Engineering journal	RE	97
Software Practice and Experience	SPE	7
Software Quality Journal	SQJ	6
ACM Transactions on Software Engineering and Methodology	TOSEM	8
IEEE Transactions on Software Engineering	TSE	53
Journal of the ACM		0
Journal of Software Maintenance and Evolution: Research and Practice		0
IEEE Transactions on Dependable and Secure Computing		0
Sum		**349**

5.1 Publication and Selection Bias

Including only papers from journals classified as by ISI as belonging to software engineering limits the possibility to generalize the results to other forums in which requirements engineering technologies are published. This also introduces the risk of missing technologies and evaluations published in conference proceedings, technical reports, workshops etc. However, as the major journals in the software engineering field are included in the review, this threat should be limited.

The selection of papers from journals is also a threat to validity. First, the search procedure used introduces a threat as it could miss papers relevant for inclusion in the review. However, the search terms used did not miss relevant papers when applied to REj (Ivarsson and Gorschek 2009). Even if the search missed papers, it should not introduce any systematic bias with respect to the results. Second, the inclusion criterion is first applied by reading the abstract of the papers. This introduces a threat as the abstract does not necessarily reflect what is actually presented in the papers. This threat was investigated in REj and found to be limited (Ivarsson and Gorschek 2009).

5.2 Data Extraction

A potential threat to validity is the subjective judgment used to include/exclude papers and to extract data from the papers that were included. To limit this threat, a pilot trial was

carried out with the classification scheme and inclusion criteria, and these were changed prior to use as described in Section 7.1. In addition, the aspects and scoring rubrics used for data extraction are derived from the research questions. With respect to the actual data extraction, the scoring of aspects is subject to subjective variations. For instance, there maybe several plausible classifications for one paper. To limit this threat, papers were classified giving researchers the benefit of the doubt, i.e. papers were classified in accordance with what is mentioned in the papers. In addition, when several classifications are possible, the one most beneficial for the subsequent analysis was chosen. This means that the results presented in this review are in some sense a best case scenario, e.g. example applications by the researcher himself/herself are sometimes reported as "case studies", giving a higher relevance score.

6 Results from Using the Model

This section presents the findings of the review and is arranged according to the research questions presented in Section 4.1. The quantification and visualization techniques from the model described in Section 3 are used to analyze the data from the review.

6.1 RQ1: What is the State-of-Practice in Relation to Rigor and Relevance of Technology Evaluations in Requirements Engineering?

The variables for rigor and relevance found in the articles included in the review are given in Fig. 1. The size of the bubbles indicates the number of papers in each class. The maximum value for rigor a paper can have is three, while relevance has a maximum of four. The figure shows that the majority of evaluations end up in the lower left quadrant of the bubble chart, indicating a lack of both rigor and relevance. 116 articles have zero rigor and relevance. This means that about one third of all the evaluations included in this review are experiments in which aspects related to rigor are not described or are examples of application of a technology done by either students or researchers in academia in toy examples. This is disappointing from an academic perspective as it is difficult to synthesize and gain a better understanding of how technologies actually perform if there is no actual evaluation or reporting of evaluations. Example applications often do not evaluate the usefulness or usability of a technology but rather simply illustrate that it can be applied to an example. The concentration of evaluations ending up in the left half of the chart is also disappointing from a technology transfer perspective, as these evaluations have less potential for actually influencing practice. Furthermore, few evaluations landing in the upper part of the chart illustrate that aspects needed to evaluate the quality (rigor) of evaluation are often lacking in the reporting. This can hamper the progress of research as, even if interesting results are published, the possibility to further investigate the issue through replication or reproduction can be limited.

The papers that end up having the highest levels of rigor and relevance are given in Table 10.

6.2 RQ2: Have Rigor and Relevance Changed Over Time?

To analyze the evolution of research, Fig. 2 shows how the variables of rigor and relevance have changed over time together with the number of papers included each year. Rigor and relevance are scored on the left y-axis while the number of papers is given on the right y-axis.

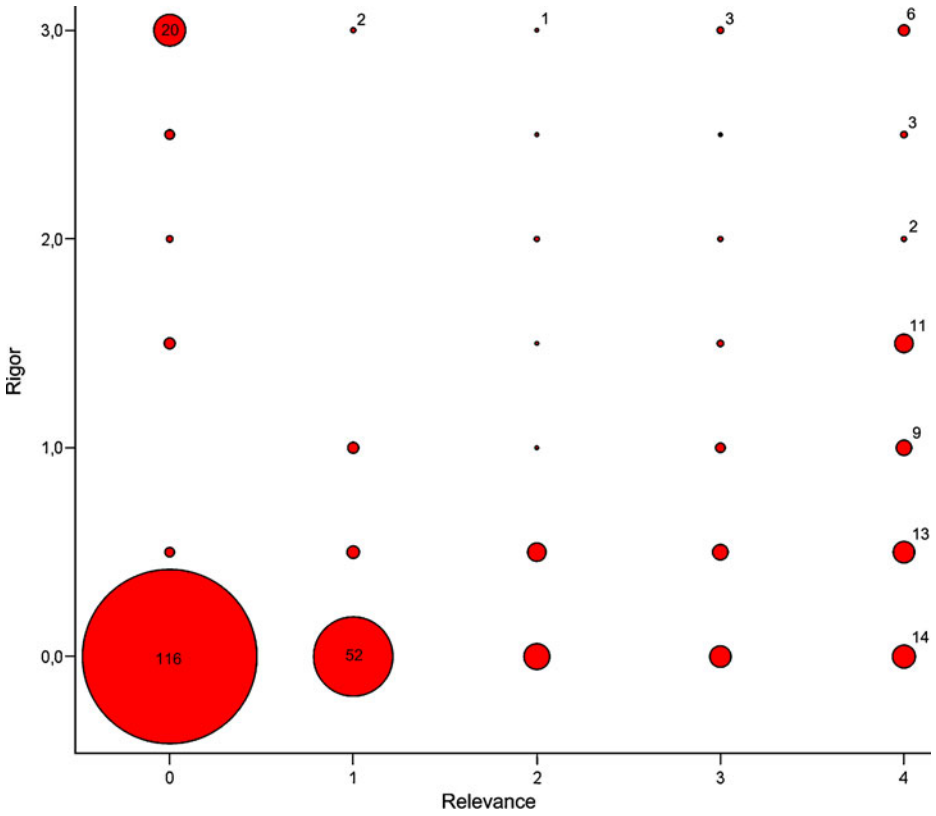

Fig. 1 Rigor and relevance of evaluations

Rigor has a maximum value of three, and relevance can be at most four. The results indicate an improvement in both rigor and relevance. Rigor matures from a rating of around 0.5 during most of the 1990s to about one in the 21st century. This is a doubling over 10 years, which is an improvement, even though the average rigor is still not high (on average, only one of the three aspects is described).

With regard to relevance, the increasing number of papers that present technology evaluations supports the results given in (Zelkowitz 2009), which shows increasing levels of validation. However, the industrial relevance or realism of evaluations does not seem to

Table 10 Papers with highest level of rigor and relevance

		Relevance	
		3	4
Rigor	**3**	(Carlshamre 2002), (Arthur and Gröner 2005), (Berling and Runeson 2003)	(Regnell et al. 2001), (Lauesen and Vinter 2001), (Gorschek et al. 2007), (Jiang et al. 2008), (Anda et al. 2006), (Karlsson et al. 2006)
	2,5	(Mich et al. 2005)	(Maiden et al. 2005), (Emam and Madhavji 1996), (Laitenberger et al. 2002)
	2	(Beecham et al. 2005), (Dzida et al. 1978)	(Hall et al. 2002), (Aurum et al. 2006)

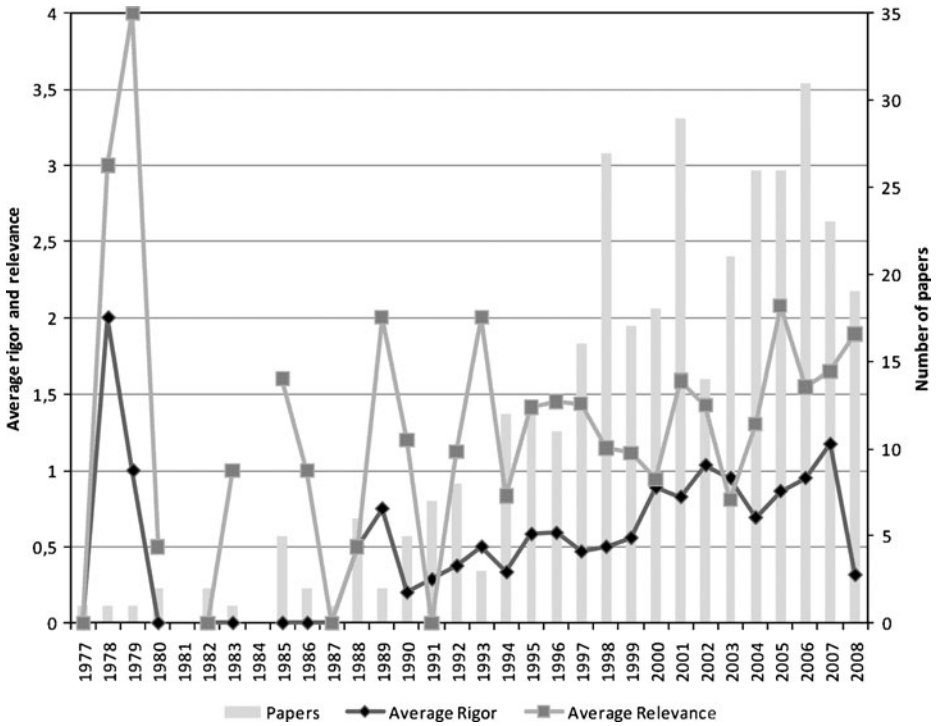

Fig. 2 Average rigor and relevance over time

show any noticeable improvement over time (even though the most recent years indicate a slight improvement). Thus, while the results show an increasing number of potentially relevant studies, the average relevance of research papers remains about the same.

6.3 RQ3: Does Publication Venue Influence Rigor and Relevance?

To investigate the influence that publication venue has on articles, the average rigor and relevance of each journal are calculated. This is illustrated in Fig. 3. The number of papers included in each journal is indicated on the right y-axis. What is apparent from the results is that the rigor and relevance of articles varies between different publication venues. The Empirical Software Engineering (EMSE) journal stands out in terms of rigor. No other journal comes near the same level of rigor. The Software Quality journal (SQJ) has the highest relevance score followed by IEEE Software and Computer. On the other side of the spectrum, it can be seen that the flagship journals in software engineering, Transactions on Software Engineering (TSE) and Transactions of Software Engineering and Methodology (TOSEM), surprisingly has both low relevance and rigor. This may depend on incorporating work that is more theoretical. However, pure theoretical work is not included in this review as the simplest form of evaluation is chosen to be a technology applied to an example. Furthermore, journals with page limitations, e.g. IEEE Software and Computer, show low levels of rigor. When the number of pages is limited, the first thing to go seems to be describing aspects related to rigor. This maybe good enough for these journals if the results are more thoroughly described elsewhere.

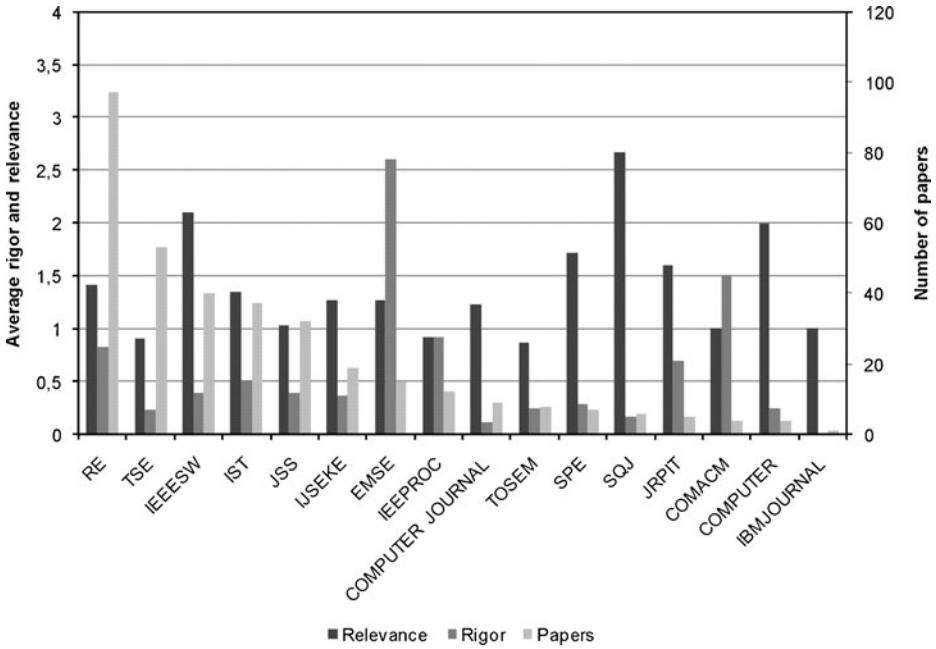

Fig. 3 Average rigor and relevance of each journal

7 Validation of the Model

The review presented in Sections 4 and 6 provides a case for applicability of the model to technology evaluations in requirements engineering. This section gives an account of how the model was constructed and validated and how the results of applying the model can be used. Section 7.1 describes how in the review the model evolved through usage. Section 7.2 discusses the scoring of research and potential implications. Section 7.3 discusses how the results of the systematic review, and thus the model, can be used.

7.1 Model Evolution

The systematic review acted as a validation of the model, showing that the model could be applied, and doubled as a case on which the model was refined. The model thus evolved based on lessons learned from the systematic review to improve its ability to characterize research. The interesting thing is that the evolution of the model actually consisted of reducing the model (making it more basic and forgiving). This was necessary, as the majority of the technology evaluations (studies) were lacking in any advanced characterizations other than the simple ones seen in the systematic review in Section 6. The model should relate to state-of-research and only include aspects that several studies/papers mention to avoid ending up with many detailed aspects applicable to only a very few studies (if any). Thus, the model used in the systematic review is actually the simplified version of the initially designed model. The simple model came out of the necessity to offer a more forgiving and less detailed evaluation of present research. This is confirmed if we look at the results of the systematic review carried out with the simplified model. Looking at the classifications uncovered in the review (see Fig. 1), about 10% of the papers have the

highest level of rigor and 15% have the highest level of relevance. This means that the quite basic model relates rather well to state-of-research. Having more detailed aspects would only offer even fewer (if any) papers classified at higher levels. The original model (prior to adapting it to reality and simplifying it) included additional aspects. These can be seen in Table 11.

Source of problem statement was considered for inclusion in the classification. If the problem tackled in an evaluation originates from industry, the results are presumably more relevant than if an academic problem is investigated. This aspect was removed, however, since very few papers mentioned having an empirical basis or industry need that motivated the research that was carried out. In addition, not reporting industrial motivation might be a style of writing, i.e. research is motivated in relation to previous research.

Describing the *expertise and skill of subjects* in relation to the technology under evaluation is important in terms of indicating the population for which the results holds true (Sjøberg et al. 2007). Alas, the expertise of subjects used in evaluations is seldom mentioned in papers. Instead, a simplified classification of subjects is included in the classification, identifying only whether the subjects are researchers, students or professionals.

Different aspects can be considered in classifying the scale of evaluations, including *task duration* and the scale of the application. The duration of tasks performed in evaluations is seldom mentioned in papers. Instead, only the scale of the application is included in the model, as this is often reported.

The way in which the study is carried out in terms of what *dependant variables* are measured is not considered in the final model. The variables measured should be relevant to practitioners for the results to be relevant for industry (Jedlitschka et al. 2007). This aspect was excluded from the classification as it is difficult to incorporate the aspect into an overall scoring of rigor and relevance. It is also difficult to discern what constitutes a relevant dependant variable in requirements engineering (Gorschek and Davis 2008), and the relevance of variables measured is likely also dependant on the technology investigated.

Table 11 Aspect excluded from the model

Aspect	Description	References
Relevance		
Source of problem statement	What is the source of the problem addressed by the technology? If the technology is developed to address needs in industry it is more likely to be relevant for industry.	(Khurum and Gorschek 2009)
Expertise and skill of subjects	What is the skill and expertise of subjects used in the evaluation in relation to the technology being evaluated?	(Sjøberg et al. 2005)
Task duration	What is the duration of the evaluation?	(Sjøberg et al. 2005)
Dependant variables considered (efficiency, effectiveness, quality)	Are dynamic properties of technology usage measured or is it is only a case of application?	
Rigor		
Conclusion validity/Reliability	Is the conclusion validity or reliability of the study discussed?	(Wohlin et al. 2000)
External validity	Is the external validity discussed?	(Wohlin et al. 2000)
Internal validity	Is the internal validity discussed?	(Wohlin et al. 2000)
Construct validity	Is the construct validity discussed?	(Wohlin et al. 2000)

A decision was also made to exclude more fine-grained aspects concerning the validity of the evaluation, as validity is seldom reported in papers. Thus, it was decided to include only one aspect, comprising all parts, for classifying validity.

7.2 Scoring of Research

With respect to the scoring rubrics used to classify studies and the quantification of the classification into variables, it is important that the valuation actually approximates the relevance and rigor of studies. Rigor is scored to reflect the extent to which aspects related to rigor are reported. The scoring should enable locating studies described to an extent that facilitates replication, reproducing and synthesis of evidence. In this respect, the quantification is straightforward in assigning higher values to studies that described aspects to a greater extent and should thus provide a good approximation of what studies lend themselves to replication, reproduction and synthesis.

The scoring for relevance should reflect the potential for evidence to impact industry. Evidence obtained in realistic settings is more likely to facilitate generalization to the target environment (industry). Thus, when valuing context, subjects and scale, a higher score is assigned when the aspects are classified as being realistic. The research method used in studies also influences how compelling evidence is to practitioners. Zelkowitz et al. (1998) found that research methods relevant for practitioners' environments, e.g. case studies, are valued higher than, for example, experiments. This can in part depend on these research methods lending themselves to investigations of real world situations. However, it can also depend on the state-of-research, i.e. the realism of experiments is usually not high (Sjøberg et al. 2005). Table 12 shows the distribution of context, scale and subjects for different research methods as found in the systematic review presented in Sections 4 and 6. Research methods that are scored as contributing to relevance in the model presented in Section 3 often have context, subjects and scale that are realistic. The majority of studies reported as lessons learned, field studies, surveys and interviews have realistic scale, subjects and context. Action research studies are also often conducted in realistic settings but use researchers as subjects. This can be expected, however, as, in action research studies, researchers sometimes act as experts/investigators that drive change (Robson 2002). However, case studies show a mix of realistic and unrealistic traits. It is often emphasized in the literature that case studies per definition are conducted in real world settings (Yin 2008;

Table 12 Context, subjects and scale used with different research methods

Research method	Context		Subjects				Scale			
	Academia	Industry	Practitioner	Researcher	Student	Not mentioned	Toy	Down scaled	Industrial	Not mentioned
Action research	1	8	3	6	0	0	1	0	8	0
Example application	140	10	2	147	1	0	108	13	29	0
Lessons learned	3	14	11	6	0	0	0	0	16	1
Case study	50	58	36	66	5	1	29	5	74	0
Field study	0	4	4	0	0	0	0	0	4	0
Experiment (HS)	36	7	8	2	33	0	39	0	4	0
Experiment (SW)	9	2	0	11	0	0	3	0	8	0
Survey	0	4	4	0	0	0	0	0	2	2
Interview	0	3	3	0	0	0	1	0	2	0

Runeson and Höst 2009). This is not reflected in the results of the review and is most likely due to a large misuse of the term "case study" in the papers reviewed.

Looking at experiments using both software and human subjects, these are seldom carried out in a realistic setting. This maybe one reason for practitioners to value case studies and lessons over experiments. Considering the state-of-research presented in Table 12, the valuation of research methods used in the model presented in Section 3 seems to approximate the actual relevance. Given the distribution, studies ending up with a relevance score of three or four should be interpreted as relevant, as this includes action research studies performed with researchers as subjects and experiments carried out in a realistic setting.

The academic impact or relevance, in terms of average number of citations per year for the different levels of industry relevance,[1] is given in Table 13. There is no indication that the level of industry relevance, in the model presented in this paper, influence academic relevance.

7.3 Using the Model

Researchers can use the 10% of studies with the highest level of rigor (see Fig. 1) to synthesize evidence in the field. This is possible because the studies are described in such a degree that the evidence can be scrutinized and judged. These studies are also prime candidates for replication or reproduction. Researchers looking for motivation on the part of industry for studies or practitioners searching for evidence of usability and usefulness of technologies in an industrial setting can use the 15% of studies rated as having the highest industrial relevance (see Fig. 1).

However, the results given in Section 6 show that the majority of evaluations lack both rigor and relevance. To increase the number of studies that can be used for the above mentioned objectives, the results presented in Section 6 emphasize ways to improve the state-of-research in terms of rigor and relevance. Figure 4 illustrates three different alternatives to increasing rigor and relevance of evaluations in Fig. 1. In planning evaluations, researchers can use these alternatives to ensure an adequate rigor and relevance of studies. Each ways are discussed below.

A) Increasing Rigor of Articles with Low Rigor and Industrial Relevance

It has been argued that rigor is irrelevant until relevance has been established (Keen 1991), suggesting that the most important way to improve evaluations is to focus on Arrow B in Fig. 4. This is not the exact standpoint taken in this paper. There is still a necessity to experiment and investigate the underlying assumptions of technologies. The main thing is to keep evaluations relevant. This means that two issues must be considered: the topic investigated and how evaluations are performed. Distinguishing between relevant and non-relevant evaluations is not easy, as there is no way of knowing this up front. There is also no guarantee that academic relevance overlaps industrial relevance. Earlier literature reviews have found that topics studied empirically are quite narrow (Höfer and Tichy 2007). To improve the situation, studies with less industrial relevance should offer academic value in terms of rigor by shifting the focus from application examples performed by the researcher him/herself to actual evaluations using experiments, for example. Experiments have a higher potential to enable an understanding of the technology under evaluation and facilitate

[1] The number of citations was retrieved on August 27, 2010, using Scopus. Eighty-eight percent of the included papers were indexed in Scopus.

Table 13 Average number of citations per year for the different levels of relevance

Relevance	Number of papers	Average number of citations per year
4	54	2,58
3	32	1,88
2	29	1,76
1	57	2,88
0	136	1,91

synthesizing evidence. Further, conducting more experiments would potentially overcome the problem of a narrow focus of empirical research. If more experiments are performed, it is likely that the scope of what is investigated will broaden.

B) Increasing the Relevance of Articles

The most significant way to aid technology transfer seems to be to promote research relevant to industry. This not only improves the technology transfer decision support offered to practitioners but also enables feedback from industry to academia, i.e. evaluations in industry can provide feedback to academia on what works in practice. This constitutes one way of improving the relevance of technologies developed in academia.

The first thing to do to increase the relevance of technology evaluations is to perform them in settings similar to ones in which they are intended to be used. This implies having a realistic scale, subjects and context in evaluations. In addition to the setting, the way in which evaluations are performed influences their realism. Using off-line experiments, it is difficult to emulate real usage with respect to scale and the duration of tasks. This calls for dynamic evaluations where technologies are used in real software development in which they are monitored and evaluated. However, a prerequisite for performing dynamic evaluations in industry is that the basics of the technology are known. Dynamic evaluations are meant to address scaling issues and sort out teething problems, e.g. tailoring technology for practical use. Thus, failing to emphasize an understanding of the basics of a technology beforehand could render dynamic

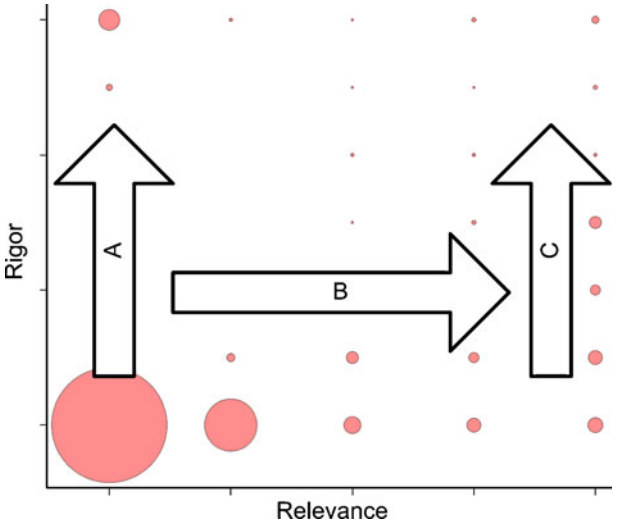

Fig. 4 Improving state-of-research

evaluations useless. Researchers need to pursue an iteratively improved understanding of the technologies under evaluation (illustrated in Fig. 5). Going directly for dynamic evaluations in industry is risky as there might be numerous issues that are not understood that would jeopardize the effort. Static evaluations, i.e., off-line small-scale evaluations in the form of experiment, workshops or pilot applications, in either academia or industry, are needed to understand the basics of the technology, e.g. the efficiency and effectiveness of a technology. Pilot evaluations in industry are used to tailor and collect initial feedback on practical issues in the particular technology using limited resources. Then, when the technology is understood and tailored for practical use, dynamic evaluations utilizing case studies, action research etc. can be pursued. This also enables continuous feedback to technology development in academia, as industry feedback adds valuable information beyond pure academic feedback.

In addition, getting companies to commit to evaluations alone indicates the relevance of technologies. It is unlikely to that commitment will be gotten to an idea that is not perceived as important. If research is pursued in this way, bad ideas are likely to be discarded more quickly (Glass 1994). This also limits the risk of discarding good ideas that are simply not yet mature as they are iteratively tested and refined.

C) Increasing Rigor of Relevant Articles

In the case of studies that have a high potential for impacting industry, e.g. studies that show improvements in industry, failing to report adequately lessens their value. From an academic perspective, failing to report study design and validity makes it difficult to evaluate the quality of the evaluation. Empirical work is not perfect in contrast to theoretical work, and failing to report aspects of the study makes evaluating them a question of guesswork. It also limits the possibilities to replicate or reproduce the study and thus strengthen the evidence. From an industrial standpoint, failing to report the context of the evaluation makes it hard for practitioners to understand whether the results are valid for them.

8 Future Work

This review considers only evaluative research in requirements engineering with a focus on work published in the included journals. In order to compare and contrast different sub-fields in e.g. software engineering, we encourage other researchers to conduct similar reviews with a broader or different scope. The model used in this review might need to be

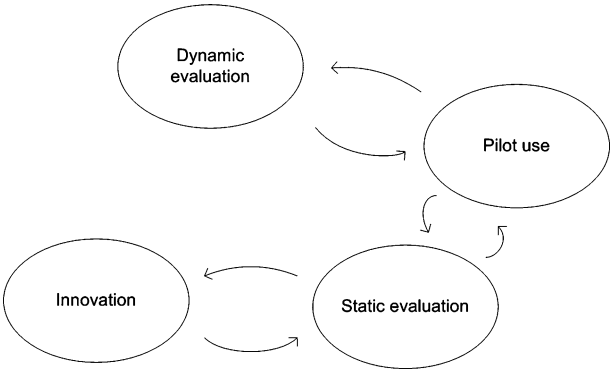

Fig. 5 Step-wise evaluation of technologies (adapted from (Gorschek et al. 2006))

changed to reflect research in specific areas, i.e., different research areas might have different levels of validation and the model should be changed to reflect state-of-research in the area. While the model can and should be adapted to fit specific subject areas, it should also evolve towards a more precise and demanding model as research design, execution and reporting improve.

9 Conclusions

This paper presents a model for evaluating the rigor and industrial relevance of technology evaluations, and introduces the use of candidate rubrics for this purpose. The model can be used by practitioners to gauge the relevance and validity of available evidence, and by researchers to point out evidence that can be synthesized, or give information that would enable replication or reproduction. Researchers get valid results they can build on, and industry professionals get decision support in terms of giving motivation to try out research results.

The model is validated through a systematic literature review that investigates the rigor and industrial relevance of technology evaluations in requirements engineering. The aim is to show the applicability of the model and to characterize how evaluations are carried out and reported to evaluate the state-of-research. The major findings of the review are;

- The number of evaluations presented in requirements engineering is increasing.
- The majority of technology evaluations in requirements engineering lacks both industrial relevance and rigor. This means that most evaluations are performed in unrealistic environments and that the reporting lacks descriptions of study design, context and validity.
- A very small fraction of evaluations (six out of 349) exhibits the highest level of industrial relevance and rigor.
- The research field does not show any improvement in terms of average industrial relevance over time. Even though some excellent evaluations are carried out, the field as such seems to maintain a *status quo* in terms of relevance. While the average rigor of evaluations has doubled in the last 10 years, it is still at a very low level. However, as evaluations are becoming more commonplace, more relevant and rigorous evaluations are being performed.
- There is a significant difference between publication venues when it comes to industrial relevance and rigor. This might indicate that reviewing policies and/or differences in quality demands impact research reporting.

The results presented pertain only to technology evaluations in requirements engineering and especially the journals included in this review. However, as the journals are the premier publication venues for researchers in requirements engineering, they are likely to contain the most mature research in the field.Regarding the relevance of research carried out, different indexes related to number of publications and citations are currently often used to assess the relevance or impact of research. The results of this review stress the need for a more balanced view of relevance in software engineering research. Number of publications and citations does not consider the potential for research to impact industry, which is the most important factor determining success in an applied research field such as software engineering (Sjøberg et al. 2002). In this respect, the model presented in this paper can provide a more balanced view of relevance of research. The model can also be used preemptively in order to improve state-of-research. The model emphasizes aspects that are important in planning and reporting research and can thus be used by researchers when they plan and report studies.

References

Afzal W, Torkar R et al (2008) A systematic mapping study on non-functional search-based software testing. 20th International Conference on Software Engineering and Knowledge Engineering (SEKE 2008)

Afzal W, Torkar R et al (2009) A systematic review of search-based testing for non-functional system properties. Inf Softw Technol 51(6):957–976

Anda B, Hansen K et al (2006) Experiences from introducing UML-based development in a large safety-critical project. Empir Softw Eng 11(4):555–581

Arisholm E, Sjoberg DIK (2004) Evaluating the effect of a delegated versus centralized control style on the maintainability of object-oriented software. IEEE Trans Softw Eng 30(8):521–534

Arisholm E, Gallis H et al (2007) Evaluating pair programming with respect to system complexity and programmer expertise. IEEE Trans Softw Eng 33(2):65–86

Arthur JD, Gröner MK (2005) An operational model for structuring the requirements generation process. Requir Eng 10(1):45

Aurum A, Wohlin C et al (2006) Aligning software project decisions: a case study. Int J Software Engineer Knowledge Engineer 16(6):795–818

Basili VR, Selby RW et al (1986) Experimentation in software engineering. IEEE Trans Softw Eng 12 (7):733–743

Beecham S, Hall T et al (2005) Using an expert panel to validate a requirements process improvement model. J Syst Softw 76(3):251–275

Benbasat I, Zmud RW (1999) Empirical research in information systems: the practice of relevance. MIS Quart 23(1):3–16

Berling T, Runeson P (2003) Evaluation of a perspective based review method applied in an industrial setting. IEE Proc Softw 150(3):177–184

Carlshamre P (2002) Release planning in market-driven software product development: provoking an understanding. Requir Eng 7(3):139

Dybå T, Kampenes VB et al (2006) A systematic review of statistical power in software engineering experiments. Inf Softw Technol 48(8):745–755

Dybå T, Dingsøyr T et al (2007) Applying systematic reviews to diverse study types: an experience report. First International Symposium on Empirical Software Engineering and Measurement (ESEM)

Dzida W, Herda S et al (1978) User-perceived quality of interactive systems. IEEE Trans Softw Eng 4(4):270–276

Emam K, Madhavji NH (1996) An instrument for measuring the success of the requirements engineering process in information systems development. Empir Softw Eng 1(3):201–240

Feldt R, Höst M et al (2009) Generic skills in software engineering master thesis projects: towards rubric-based evaluation. 22nd Conference on Software Engineering Education and Training

Glass RL (1994) The software-research crisis. IEEE Softw 11(6):42–47

Glass RL (1997) Pilot studies: what, why and how. J Syst Softw 36(1):85–97

Glass RL, Vessey I et al (2002) Research in software engineering: an analysis of the literature. Inf Softw Technol 44(8):491–506

Gorschek T, Davis AM (2008) Requirements engineering: In search of the dependent variables. Inf Softw Technol 50(1–2):67–75

Gorschek T, Garre P et al (2006) A model for technology transfer in practice. IEEE Softw 23(6):88–95

Gorschek T, Garre P et al (2007) Industry evaluation of the requirements abstraction model. Requir Eng 12 (3):163–190

Hall T, Beecham S et al (2002) Requirements problems in twelve software companies: an empirical analysis. IEE Proc Softw 149(5):153–160

Hannay JE, Sjoberg DIK et al (2007) A systematic review of theory use in software engineering experiments. IEEE Trans Softw Eng 33(2):87–107

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569–16572

Höfer A, Tichy W (2007) Status of empirical research in software engineering. In: Basili V, Rombach D et al (ed) Empirical software engineering issues. Critical Assessment and Future Directions, LNCS, vol 4336 Springer, Berlin/Heidelberg, pp 10–19

Hsia P, Davis AM et al (1993) Status report: requirements engineering. IEEE Softw 10(6):75–79

IEEE (1990) IEEE standard glossary of software engineering terminology. IEEE Std 610:12–1990

Ivarsson M, Gorschek T (2009) Technology transfer decision support in requirements engineering research: a systematic review of REj. Requir Eng 14(3):155–175

Jedlitschka A, Ciolkowski M et al (2007) Relevant information sources for successful technology transfer: a survey using inspections as an example. First International Symposium on Empirical Software Engineering and Measurement (ESEM)
Jiang L, Eberlein A et al (2008) A case study validation of a knowledge-based approach for the selection of requirements engineering techniques. Requir Eng 13(2):117–146
Juristo N, Moreno AM et al (2002) Is the European industry moving toward solving requirements engineering problems? IEEE Softw 19(6):70–77
Kaindl H, Brinkkemper S et al (2002) Requirements engineering and technology transfer: obstacles, incentives and improvement agenda. Requir Eng 7(3):113–123
Kampenes VB, Dybå T et al (2007) A systematic review of effect size in software engineering experiments. Inf Softw Technol 49(11–12):1073–1086
Karlsson L, Regnell B et al (2006) Case studies in process improvement through retrospective analysis of release planning decisions. Int J Software Engineer Knowledge Engineer 16(6):885–915
Keen PGW (1991) Relevance and rigor in information systems research: Improving quality, confidence, cohesion and impact. In: H.-E. Nissen, H. Klein and R. Hirschheim (ed) Information Systems Research: Contemporary Approaches & Emergent Traditions. Elsevier, Amsterdam 27–49
Khurum M, Gorschek T (2009) A systematic review of domain analysis solutions for product lines. J Syst Softw 82(12):1982–2003
Kitchenham BA (1996–1998) Evaluating software engineering methods and tools, Part 1 to 12. ACM SIGSOFT Software Engineering Notes 21–23
Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University Joint Report
Kitchenham BA, Pfleeger SL (2001–2003) Principles of survey research, Part 1 to 6. ACM SIGSOFT Software Engineering Notes 26–28
Kitchenham BA, Pfleeger SL et al (2002) Preliminary guidelines for empirical research in software engineering. IEEE Trans Softw Eng 28(8):721–734
Laitenberger O, Beil T et al (2002) An industrial case study to examine a non-traditional inspection implementation for requirements specifications. Empir Softw Eng 7(4):345–374
Lau F (1999) Toward a framework for action research in information systems studies. Inf Technol People 12 (2):148–176
Lauesen S, Vinter O (2001) Preventing requirement defects: an experiment in process improvement. Requir Eng 6(1):37–50
Maiden N, Manning S et al (2005) Generating requirements from systems models using patterns: a case study. Requir Eng 10(4):276–288
Meyer B, Choppy C et al (2009) Viewpoint: research evaluation for computer science. Commun ACM 52 (4):31–34
Mich L, Anesi C et al (2005) Applying a pragmatics-based creativity-fostering technique to requirements elicitation. Requir Eng 10(4):262
Morris P, Masera M et al (1998) Requirements engineering and industrial uptake. Requir Eng 3(2):79–83
Moskal BM (2000) Scoring rubrics: what, when and how. Pract Assess Res Eval 7(3). http://PAREonline.net/getvn.asp?v=7&n=3. Accessed 30 September 2010
Neill CJ, Laplante PA (2003) Requirements engineering: the state of the practice. IEEE Softw 20(6):40–45
Parnas DL (2007) Stop the numbers game. Commun ACM 50(11):19–21
Perry DE, Porter AA et al (2000) Empirical studies of software engineering: a roadmap. International Conference on Software Engineering (ICSE)
Petersen K, Wohlin C (2009) Context in industrial software engineering research. Proceedings 3rd International Symposium on Empirical Software Engineering and Measurement, Orlando, USA
Petersen K, Feldt R et al (2008) Systematic mapping studies in software engineering. 12th International Conference on Evaluation and Assessment in Software Engineering, Bari, Italy
Pfleeger SL (1994–1995) Experimental design and analysis in software engineering, Parts 1 to 5. ACM SIGSOFT Software Engineering Notes 19–20
Pfleeger SL (1999) Understanding and improving technology transfer in software engineering. J Syst Softw 47(2–3):111–124
Pfleeger SL, Menezes W (2000) Marketing technology to software practitioners. IEEE Softw 17(1):27–33
Potts C (1993) Software-engineering research revisited. IEEE Softw 10(5):19–28
Redwine ST, Riddle WE (1985) Software technology maturation. 8th international conference on Software engineering, London, England, IEEE Computer Society Press
Regnell B, Höst M et al (2001) An industrial case study on distributed prioritisation in market-driven requirements engineering for packaged software. Requir Eng 6(1):51–62

Robson C (2002) Real world research. Blackwell Publishing, Cornwall
Ross DT, Schoman KE Jr (1977) Structured analysis for requirements definition. IEEE Trans Softw Eng 3(1):6–15
Runeson P, Höst M (2009) Guidelines for conducting and reporting case study research in software engineering. Empir Softw Eng 14(2):131–164
Sjøberg DIK, Anda B et al (2002) Conducting realistic experiments in software engineering. 18th International Symposium on Empirical Software Engineering (ISESE)
Sjøberg DIK, Hannay JE et al (2005) A survey of controlled experiments in software engineering. IEEE Trans Softw Eng 31(9):733–753
Sjøberg DIK, Dybå T et al (2007) The future of empirical methods in software engineering research. Future of Software Engineering (FOSE)
Šmite D, Wohlin C et al (2010) Empirical evidence in global software engineering: a systematic review. Empir Softw Eng. doi:10.1007/s10664-009-9123-y
Tichy WF, Lukowicz P et al (1995) Experimental evaluation in computer science: a quantitative study. J Syst Softw 28(1):9–18
Wieringa R, Heerkens J (2006) The methodological soundness of requirements engineering papers: a conceptual framework and two case studies. Requir Eng 11(4):295–307
Wohlin C (2009a) An analysis of the most cited articles in software engineering journals—2002. Inf Softw Technol 51(1):2–6
Wohlin C (2009b) A new index for the citation curve of researchers. Scientometrics 81(2):521–533
Wohlin C, Runeson P et al (2000) Experimentation in software engineering. Kluwer Academic Publishers, Boston
Wong WE, Tse TH et al (2009) An assessment of systems and software engineering scholars and institutions (2002–2006). J Syst Softw 82(8):1370–1373
Yin RK (2008) Case study research: Design and methods. Sage publications, Beverly Hills
Zannier C, Melnik G et al (2006) On the success of empirical studies in the international conference on software engineering. Proceedings of the 28th International Conference on Software Engineering. Shanghai, China, ACM
Zelkowitz MV (2009) An update to experimental models for validating computer technology. J Syst Softw 82(3):373–376
Zelkowitz MV, Wallace D (1997) Experimental validation in software engineering. Inf Softw Technol 39 (11):735–743
Zelkowitz MV, Wallace DR et al (1998) Culture conflicts in software engineering technology transfer. Maryland, University of Maryland, College Park

Martin Ivarsson is a graduate student at the School of Computer Science and Engineering at Chalmers University of Technology, Sweden. His research interests include empirical software engineering and software process improvement. Ivarsson received an MSc in computer engineering from Chalmers University of Technology in 2004.

Tony Gorschek is an associate professor of software engineering at the Blekinge Institute of Technology, Sweden. He also works as a consultant and manages his own company. His research interests include requirements engineering, technical product management, process assessment and improvement, and quality assurance. He conducts most of his research in close collaboration with industry. Gorschek received a PhD in software engineering from the Blekinge Institute of Technology. He is a member of the IEEE. Contact him at Blekinge Inst. of Technology, PO Box 520, SE-372 25 Ronneby, Sweden; Tony.Gorschek@bth.se