The Journal of Systems and Software 97 (2014) 65–85

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

A systematic literature review on the industrial use of software process simulation

Nauman Bin Ali*, Kai Petersen, Claes Wohlin

Blekinge Institute of Technology, Karlskrona, Sweden

ARTICLE INFO

Article history:
Received 3 February 2014
Received in revised form 29 June 2014
Accepted 30 June 2014
Available online 9 July 2014

Keywords:
Software process simulation
Systematic literature review
Evidence based software engineering

ABSTRACT

Context: Software process simulation modelling (SPSM) captures the dynamic behaviour and uncertainty in the software process. Existing literature has conflicting claims about its practical usefulness: SPSM is useful and has an industrial impact; SPSM is useful and has no industrial impact yet; SPSM is not useful and has little potential for industry.
Objective: To assess the conflicting standpoints on the usefulness of SPSM.
Method: A systematic literature review was performed to identify, assess and aggregate empirical evidence on the usefulness of SPSM.
Results: In the primary studies, to date, the persistent trend is that of proof-of-concept applications of software process simulation for various purposes (e.g. estimation, training, process improvement, etc.). They score poorly on the stated quality criteria. Also only a few studies report some initial evaluation of the simulation models for the intended purposes.
Conclusion: There is a lack of conclusive evidence to substantiate the claimed usefulness of SPSM for any of the intended purposes. A few studies that report the cost of applying simulation do not support the claim that it is an inexpensive method. Furthermore, there is a paramount need for improvement in conducting and reporting simulation studies with an emphasis on evaluation against the intended purpose.

1. Introduction

Delivering high quality software products within resource and time constraints is an important goal for the software industry. An improved development process is seen as a key to reach this goal. Both academia and industry are striving to find ways for continuous software process improvement (SPI). There are numerous SPI frameworks and methodologies available today (Basili, 1985; Team, 2006; ISO/IEC, 2003/2006), but they all have one challenge in common: the cost of experimenting with the process change. It is widely claimed that software process simulation modelling (SPSM) can help in predicting the benefits and repercussions of a process change (Ruiz et al., 2002), thus, enabling organizations to make more informed decisions and reduce the likelihood of failed SPI initiatives.

Since the suggestion to use simulation modelling for understanding the software development process by McCall et al. (1979), there is considerable literature published over the last three decades in this area. There are a number of secondary studies on the subject that have scoped the research available on the topic (Kellner et al., 1999; Liu et al., 2009; Zhang et al., 2008a,b,d, 2010; França and Travassos, 2013).

From these studies, it can be seen that all SPSM purposes identified by Kellner et al. (1999) have been explored in SPSM research over the years. In terms of the scope of the simulation models, it has ranged from modelling a single phase of the lifecycle to various releases of multiple products (Zhang et al., 2008a, 2010). The following is a brief list of some of the proclaimed benefits of SPSM:

- Improved effort and cost estimation
- Improved reliability predictions
- Improved resource allocation
- Risk assessment
- Studying success factors for global software development
- Technology evaluation and adoption
- Training and learning

Such range of claimed potential benefits and reports of industrial application and impact (Zhang et al., 2011) give an impression that simulation is a panacea for problems in software engineering (SE). However, some authors have recently questioned the validity of

* Corresponding author. Tel.: +46 455385541
E-mail addresses: nauman.ali@bth.se (N.B. Ali), kai.petersen@bth.se (K. Petersen), claes.wohlin@bth.se (C. Wohlin).

http://dx.doi.org/10.1016/j.jss.2014.06.059
0164-1212/

these claims (Pfahl, 2014). Three positions can be delineated from literature on SPSM:

Claim 1: Software process simulation is useful in SE practice and has had an industrial impact (Zhang et al., 2011).
Claim 2: SPSM is useful however it is yet to have a significant industrial impact (Münch, 2012; Houston, 2012; Birkhölzer, 2012).
Claim 3: Questions not only the usefulness but also the likelihood and potential of being useful for the software industry (Pfahl, 2014).

In this study, we aim to aggregate and evaluate, through a systematic literature review (Kitchenham and Charters, 2007), the empirical evidence on the usefulness of SPSM in real-world settings (industry and open source software development). In essence, we aim to establish which of the claims in the SPSM community can be substantiated with evidence. The main contributions of this study can be summarized as the following:

- We attempted to substantiate the claim that SPSM is an inexpensive (Melis et al., 2006; Kellner et al., 1999) mechanism to assess the likely outcome before actually committing resources for a given change in the development process (Kellner et al., 1999).
- We attempted to characterize which SPSM approaches are useful for what purpose and under which context in real-world software development (cf. Kellner et al., 1999 for a definition of purpose and scope).
- We used a systematic and documented process to identify, evaluate, and aggregate the evidence reported for the usefulness of SPSM (Kitchenham and Charters, 2007; Ivarsson and Gorschek, 2010).
- The existing secondary studies cover literature published from 1998 till December 2008. We included any literature published till December 2012 and also considered other than typical SPSM venues and found substantially more studies (a total of 87 primary studies of which 17 are published before 1998, 46 between 1998 and 2008, and 24 after 2008) that have used SPSM in a real-world software development setting than any of the existing secondary studies.
- From the existing secondary studies, we now know that many different simulation approaches *"can be applied"* and that they *"can be useful"*. However, in this study we attempt to see if there is a progression in SPSM literature and if these claims can now be substantiated.
- By following an objective, thorough and systematic approach (detailed in Section 3) the existing research on SPSM is evaluated in an objective, unbiased manner. This well-intentioned endeavour is to identify improvement opportunities to raise the quality and steer the direction of future research.

The remainder of the paper is structured as follows: Section 2 presents the related work. Section 3 explains our research methodology. Section 4 shows the characteristics of the primary studies, followed by the review results in Section 5. Section 6 discusses the results of the systematic literature review, and Section 7 concludes the paper.

2. Related work

Using the search strategy reported in Section 3.2, we identified a number of existing reviews of the SPSM literature (Kellner et al., 1999; Liu et al., 2009; Zhang et al., 2008a,b,d, 2010; França and Travassos, 2013; Bai et al., 2011). These are mostly mapping studies that provide an overview of the SPSM research. None of these studies help to assess which of the claims about the usefulness of SPSM are backed by evidence.

Kitchenham and Charters (2007) have identified criteria to assess an existing review. From these criteria, we used the detailed check-list proposed by Khan et al. (2001) and the general questions recommended by Slawson (1997). The aim was to evaluate the existing reviews on their objectives, coverage (data sources utilized, restrictions, etc.), methodology, data extraction, quality assessment, analysis, validity and reporting. The detailed criteria are available in Ali et al. (2014).

2.1. Existing reviews

The results of our assessment using these criteria (Ali et al., 2014) on the existing literature reviews in the SPSM field are presented in the following subsections.

2.1.1. Kellner et al. (1999)

Kellner et al. (1999) provide an overview of SPSM field and identify the objectives for use of simulation, scope of simulation models and provide guidance in selecting an appropriate modelling approach. They also summarize the papers from the First International Silver Falls Workshop on Software Process Simulation Modeling (ProSim'98).

Their study was published in 1999 and there is considerable new literature available on the topic. We utilize their work to explore how the research in real-world application of SPSM has used the simulation approaches for the purposes and scopes identified in their study.

2.1.2. Zhang et al. (2008a,b,d, 2010)

Zhang et al. (2008a,b,d, 2010) reported a *"two-phase"* scoping study on SPSM research. They have used six broad questions to scope (Petersen et al., 2008) the field of SPSM. Zhang et al. (2010) also acknowledge that their study *"is also a kind of mapping study"*.

In the initial phase, Zhang et al. (2008a) performed a manual search of renowned venues for SPSM literature. In the second phase (Zhang et al., 2010), it was complemented with an electronic search in IEEE, Science Direct, Springer Link and ACM, covering literature from 1998 to 2007.

The use of only one reviewer for selection, data extraction, quality assessment and study categorization is a potential threat to the validity of their studies. With such a large amount of literature that one reviewer had to go through for these two broad studies, a reviewer is highly likely to make mistakes or overlook important information (Kitchenham et al., 2012; Wohlin et al., 2013). If only one reviewer is doing the selection there is no safety net and any mistake can result in missing out a relevant article (Kitchenham, 2010). Another shortcoming, as acknowledged by the authors is the use of a less rigorous process for conducting the review (Zhang et al., 2008a), *"the main limitation of our review is that the process recommended for PhD candidates is not as rigorous as that adopted by multiple-researchers"*.

For the tasks where a second reviewer (e.g. for data extraction from 15% of the studies in the second phase (Zhang et al., 2010)) was involved, neither the inter-rater agreement nor the mechanism for resolution of disagreements is described.

2.1.3. Liu et al. (2009)

Liu et al. (2009) primarily scoped the research on software risk management using SPSM. They seek answers for five broad scoping questions but focusing on use of SPSM in software risk management. The mapping results represent the studied purposes, the scope of the modelled processes and the tools used in the primary studies.

They used the same electronic databases as Zhang et al. (2008a, 2010) for automatic search and also manually traversed the proceedings of Software Process Simulation and Modeling Workshop (ProSim) (1998–2006), International Conference on Software Process (2007–2008), Journal of Software Process Improvement and Practice (1996–2007) and special issues of Journal of Systems and Software Volume 46, Issues 2–3, 1999 and Volume 59, Issue 3, 2001.

Like the Zhang et al. (2008a, 2010) study, Liu et al. (2009) did not use the quality assessment results in the selection of studies or in the analysis. Their entire review was done by one reviewer *"One PhD student acted as the principal reviewer, who was responsible for developing the review protocol, searching and selecting primary studies, assessing the quality of primary studies, extracting and synthesizing data, and reporting the review results."*

2.1.4. Zhang et al. (2011)

Zhang et al. (2011) present an overview of software process simulation and a historical account/time-line of SPSM research, capturing who did what and when. They claim to have done some impact analysis of SPSM research based on the results of their earlier reviews (Zhang et al., 2008a,b,d, 2010). The *"case study"* reported in this article to supplement the *"impact"* analysis is at best anecdotal and is based on *"interview-styled email communications"* with Dr. Dan Houston.

Lastly, they have acknowledged this to be an initial study that needs to be extended when they say *"we are fully aware that our results are based on the examination of a limited number of cases in this initial report. The impact analysis will be extended to more application cases and reported to the community in the near future"*.

Furthermore, the following conclusions in the article (Zhang et al., 2011) are not backed by traceable evidence reported in primary studies included in their review:

- *"It is shown that research has a significant impact on practice in the area"* i.e. SPSM in practice.
- *"Anecdotal evidence exists for the successful applications of process simulation in software companies"*.
- *"The development of an initial process simulation model may be expensive. However, in the long-term, a configurable model structure and regular model maintenance or update turn out to be more cost effective"*.

2.1.5. França and Travassos (2013)

França and Travassos (2013) characterized the simulation models in terms of model type, structure, verification and validation procedures, output analysis techniques and how the results of the simulation were presented in terms of visualization. Their study is different from previously discussed reviews (in Sections 2.1.2–2.1.4) as it considers verification and validation of models, and hence has an element of judging the quality of the simulation models being investigated. Another difference that is important to note is their inclusion of all simulation studies that were related to the software engineering domain (e.g. architecture) thus covering simulation as a whole. They used Scopus, EI Compendex, Web of Science and developed their search string by defining the population, intervention, comparison, and outcome.

In their selection of studies, one reviewer conducted the selection first, his decisions were reviewed by a second reviewer and lastly the third reviewer cross-checked the selection. This increased the validity of study selection however it is prone to bias as the selection results from the first reviewer were available to the second reviewer and subsequently both the categorizations potentially biased the selection decision of the third reviewer.

The list of primary studies and the results of quality appraisal are reported in the study. They have also reported their data extraction form, but did not report on the measures undertaken to make the data extraction and classification of studies more reliable.

They extracted information about model verification and validation and how many studies conducted this activity in different ways. This provides an interesting point of comparison with our study as both studies conducted this assessment independently without knowing each others outcomes.

2.1.6. Bai et al. (2011)

Bai et al. (2011) conducted a secondary study of empirical research on software process modelling without an explicit focus on SPSM. The study has four research questions that scope the empirical research in software process modelling for:

- Research objectives
- Software process modelling techniques
- Empirical methods used for investigation
- Rigor of studies (whether research design and execution are reported)

The study used *"7 journals and 6 conference proceedings, plus other relevant studies found by searching across online digital libraries, during the period of 1988 till December 2008"*. Although the general selection criteria and data extraction form are presented, no details of the procedure for selection, extraction or quality evaluation are presented. The detailed criteria for how the rigor of studies was evaluated are also not reported. Likewise it is unclear what was the role of each reviewer in the study. Without this information it is difficult to judge whether the results are sensitive to the way the review was conducted.

Given that only 43 empirical studies are identified in their review raises some concerns about their search strategy (selection of venues, search strings, etc.). Since their study had a broader scope than ours which is including all software process modelling literature, there should have been substantially more studies.

To aggregate results they used frequency analysis in terms of how many studies investigated a specific research objective, process modelling technique, using a certain empirical method and how many described the design and execution of the study.

2.2. Our contribution

The contributions of this study in comparison to existing secondary studies can be summarized as following:

1. Given the conflicting claims about the usefulness of SPSM it was important to use a systematic methodology to ensure reliability and repeatability of the review to the extent possible. We have decided to use two reviewers and other preventive measures (discussed in detail in Sections 3.4–3.6) to minimize the threats of excluding a relevant article. These measures included using pilots, inter-rater agreement statistics and a documented process to resolve differences. With these we aimed to reduce the bias in various steps of selection, extraction and analysis of the primary studies.
2. The conflicting positions with regard to SPSM could not be resolved based on the existing secondary studies because their focus is not to identify and aggregate evidence (as discussed in Section 2.1). On the contrary, our contribution is the identification of research studies using SPSM in real-world software development followed by an attempt to evaluate and aggregate the evidence reported in them. Thus, investigating if the claims of potential benefits can be backed by evidence.
3. In theory, a systematic literature review is an exhaustive study that evaluates and interprets *"all available research"* relevant to the research question being answered (Kitchenham and

Charters, 2007). However, in practice it is a subset of the overall population that is identified and included in a study (Wohlin et al., 2013). In this study, however, we aspired to take the study population as close to the actual population. The number of studies identified in this study compared to other reviews is discussed in detail in Section 6.6. To achieve this we took following decisions:

- Search in not restricted to only typical venues of SPSM publications and includes the databases that cover computer science and SE literature.
- Lastly in the existing reviews, the potentially relevant sources for the management and business literature were not included in the search. Zhang et al. (2010) noticed that SPSM research mainly focuses on managerial interests. Therefore, it is highly probable that SPSM studies may be published outside the typical computer science and SE venues. Thus, in this literature review, we also searched for relevant literature in data sources covering these subjects. In particular, business source premier was searched that is specifically targeting business literature.
- The secondary studies by Bai et al. (2011), Liu et al. (2009) and Zhang et al. (2008a,b,d, 2010) only cover literature published between 1998 and 2008, in this systematic literature review we do not have an explicit restriction on the start date and include all literature published till December 2012. Given the noticeable trend of increasing empirical research reported in these studies (Bai et al., 2011; Liu et al., 2009) our study also contributes by aggregating the more recent SPSM literature.
- No start date was put on the search to exhaustively cover all the literature available in the selected electronic databases up till the search date (i.e. December 2012). This enabled us to identify the earliest work by McCall et al. (1979) and also include earlier work of Abdel-Hamid (1988a,b, 1989a,b, 1990, 1993) and Abdel-Hamid and Leidy (1991).

4. Other secondary reviews only scoped the existing research literature and did not highlight the lack of evaluation of claimed benefits. Overall, in a systematic and traceable manner we identify the limitations of current research in terms of reporting quality, lack of verification and validation of models, and most significantly the need to evaluate the *usefulness* of simulation for different purposes in various contexts.
5. This review by identifying the limitations in the current SPSM research has taken the first step towards improvement. The criticism of SPSM is not intended to dismiss its use, but to identify the weaknesses, raise awareness and hopefully improve SPSM research and practice. We have also provided recommendations and potential directions to overcome these limitations and perhaps improve the chances of SPSM having an impact on practice.

3. Research methodology

To identify appropriate SPSM approaches for given contexts and conditions a systematic literature review following the guidelines proposed by Kitchenham and Charters (2007) was performed. We attempted to aggregate empirical evidence regarding the application of SPSM in a real-world settings.

3.1. Review question

To assess strength of evidence for usefulness of simulation in real-world use we attempt to answer the following research question with a systematic literature review:

- ***RQ1***: What evidence has been reported that the simulation models achieve their purposes in real-world settings?

Table 1
Digital databases used in the study.

Database	Motivation
IEEE, ACM Digital and Engineering Village (Inspec and Compendex) and Science direct	For coverage of literature published in CS and SE.
Scopus, Business source premier, Web of science	For broader coverage of business and management literature along with CS, SE and related subject areas.
Google Scholar	To supplement the search results and to reduce the threats imposed by the limited search features of some databases this search engine was used.

3.2. Need for review

As a first step in our review, to identify any existing systematic reviews and to establish the necessity of a systematic review, a search in electronic databases was conducted. The keywords used for this purpose were based on the synonyms of systematic review methodology listed by de Almeida Biolchini etg al. (2007) along with *"systematic literature review"*. The search was conducted in the databases identified in Table 1, in year 2013, using the following search string with two blocks joined with a Boolean 'AND' operator:

(software AND process AND simulation) ***AND*** *("systematic review" OR "research review" OR "research synthesis" OR "research integration" OR "systematic overview" OR "systematic research synthesis" OR "integrative research review" OR "integrative review" OR "systematic literature review")*

This search string gave 47 hits in total. After removing duplicates, titles and abstracts of the remaining articles were read. This way we identified five articles that report two systematic reviews (Zhang et al., 2008a,b,d, 2010; Liu et al., 2009).

By reading the titles of articles that cite these reviews, we identified two more relevant review articles (Zhang et al., 2011; França and Travassos, 2013). In Section 2, we have already discussed in detail the limitations of these articles. We have also discussed the novel contributions of our study and how we have attempted to overcome the shortcomings in these existing reviews.

3.3. Search strategy

A conscious decision about the keywords and data-sources was made that is detailed below along with the motivation.

3.3.1. Data sources

Since, the study is focused on the simulation of software development processes, therefore it is safe to look for relevant literature in databases covering computer science (CS) and software engineering (SE). However, as the application of simulation techniques for process improvement may be published under the business related literature, e.g. organizational change, we decided to include databases of business literature as well.

3.3.2. Keywords

Starting with the research questions suitable keywords were identified using synonyms, encyclopaedia of SE (Madachy, 2002) and seminal articles in the area of simulation (Kellner et al., 1999). The following keywords were used to formulate the search strings:

- **Population**: Software process or a phase thereof. *Alternative keywords*: software project, software development process, software testing/maintenance process.

- **Intervention**: Simulation. *Alternative keywords*: simulator, simulate, dynamic model, system dynamics, state based, rule based, Petri net, queuing, scheduling.
- **Context**: Real-world. *Alternative keywords*: empirical, industry, industrial, case study, field study or observational study. Our target population was studies done in industry and we intended to capture any studies done in that context regardless of the research method used. We expected that any experiments that have been performed in industrial settings would still be identified. Yet by not explicitly including experiment as a keyword we managed to, some an extent, disregard studies in a purely academic context.
- **Outcome**: Positive or negative experience from SPSM use. Not used in the search string.

The keywords within a category were joined by using the Boolean operator OR and the three categories were joined using the Boolean operator 'AND'. This was done to target the real-world studies that report experience of applying software process simulation. The following is the resulting search string:

((software 'Proximity Op' process) OR (software 'Proximity Op' project)) ***AND*** *(simulat* OR "dynamic model" OR "system dynamic" OR "state based" OR "rule based" OR "petri net" OR "queuing" OR "scheduling")* ***AND*** *(empirical OR "case study" OR "field study" OR "observational study" OR industr*)*

The proximity operator was used to find more relevant results and yet at the same time allow variations in how different authors may refer to a software development process, e.g. software process, software testing process, etc. However, in the databases that did not correctly handle this operator we resorted to the use of Boolean operator AND instead. The exact search strings used in individual databases can be found in Ali et al. (2014).

The search in the databases (see Table 1) was restricted to title, abstract and keywords except in Google Scholar where it was only done in the title of the publications (the only other option was to search the full-text). Google Scholar is more of a search engine than a bibliographic database. Therefore, we made a trade-off in getting a broader coverage by using it without the *context* block, yet restricting the search in titles only to keep the number of hits practical for the scope of this study.

Fig. 1 provides an overview of the search results and selection procedure (discussed in detail in Section 3.4) applied in this review to identify and select primary studies.

3.4. Study selection criteria and procedure

Before the application of selection criteria (related to the topic of the review) all search results were subjected to the following generic exclusion criteria:

- Published in a non peer reviewed venue e.g. books, Masters/Ph.D. theses, keynotes, tutorials and editorials, etc.
- Not available in English language.
- A duplicate (at this stage, we did not consider a conference article's subsequent publication in a journal as duplicate this is handled later on when the full-text of the articles was read).

Where possible, we implemented this in the search strings that were executed in the electronic databases. But since many of the journals and conferences published in primary databases are covered by bibliographic databases, we had a high number of duplicates. Also, in Google Scholar we had no mechanism to keep out grey literature from the search results. Therefore, we had to do this step manually.

After this step, the remaining articles were subjected to three sets of selection criteria preliminary, basic and advanced. As the number of search results is fairly large, for practicality, the preliminary criteria were used to remove the obviously irrelevant articles.

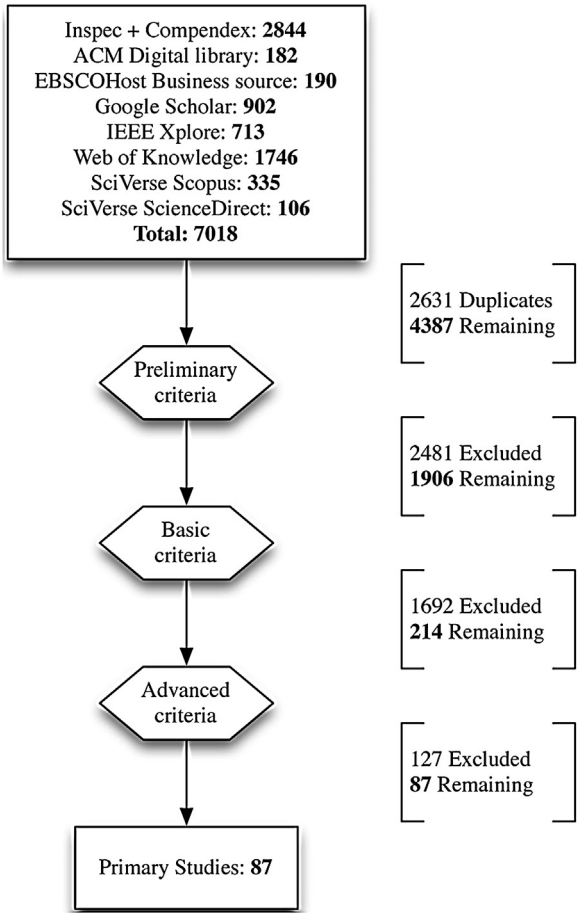

Fig. 1. Overview of steps in the selection of primary studies.

3.4.1. Preliminary criteria

Preliminary criteria were applied on the titles of the articles, the information about the venue and journal was used to supplement this decision. If the title hinted exclusion of articles but there was a doubt the abstract was read. If it was still unclear the article was included for the next step where more information from the article was read to make a more informed decision.

- Exclude any articles related to the simulation of hardware platforms.
- Exclude any articles related to the use of simulation software, e.g. simulating manufacturing or chemical process or transportation, etc.
- Exclude any articles related to use of simulation for evaluation of software or hardware reliability and performance, etc.
- Exclude any articles related to use of simulation in SE education in academia. Articles with educational focus using software process simulation were not rejected straight away based on the title. Instead we read the abstract to distinguish the SPSM used for training and education in the industry from those in a purely academic context. Only the articles in the latter category were excluded in this study. If such a decision could not be made about the context, the article was included for the next step.

The preliminary criteria were applied by only one reviewer. By using *"when in doubt, include"* as a rule of thumb we ensured inclusiveness to reduce the threat of excluding a relevant article.

Table 2
Different possible scenarios for study selection.

Reviewer 1	Reviewer 2		
	Relevant	Uncertain	Irrelevant
Relevant	A	B	D
Uncertain	B	C	E
Irrelevant	D	E	F

Also having explicit criteria about what to exclude reduced the reviewer's bias as we tried to minimize the use of authors own subjective judgement in selection.

3.4.2. Basic criteria

Basic criteria were applied to evaluate the relevance of the studies to the aims of our study by reading the titles and abstracts.

- Include an article related to the simulation of a software project, process or a phase thereof. For example, the type of articles identified by the preliminary criteria.
- Exclude an article that only presents a simulation technique, tool or approach.
- Exclude a non-industrial study (e.g. rejecting the empirical studies with students as subjects or mock data). Studies from both commercial and open source software development domains were included in this review.

It was decided that articles will be labelled as: Relevant, Irrelevant or Uncertain (if available information i.e. title and abstract, is inconclusive). Given that two reviewers will do the selection we had six possibilities (as shown in Table 2) of agreement or disagreement between the reviewers about the relevance of individual articles.

In Table 2 categories A, C and F are cases of perfect agreement between reviewers. The decision regarding each of the categories motivated by the agreement level of reviewers and likelihood of finding relevant articles in such a category is listed below:

- Articles in categories A and B (considered potential primary studies) will be directly taken to the last step of full-text reading. Although articles in category B show some disagreement between the authors but (since one author is certain about the relevance and the other is inconclusive) we considered it appropriate to include such studies for full-text reading.
- On the other hand, articles in category F will be excluded from the study as both reviewers agree on their irrelevance.
- Articles in category C will be reviewed further (by both reviewers independently using the steps of adaptive reading described below) where more detail from the article will be used to assist decision making. This was a rational choice to consult more detail, as both reviewers concurred on a lack of information to make a decision.
- Articles in categories D and E show disagreement, with category D being the worst as one author considers an article relevant and other considers it irrelevant. Articles in these two categories were deemed as candidates for discussion between reviewers. These articles were discussed and reasons for disagreement were explored. Through consensus, these articles were placed in either category A (included for full-text reading as a potential primary study), C (uncertain need more information and subjected to adaptive reading) or F (excluded from the study).

To develop a common understanding of the criteria both reviewers read the criteria and using *"think aloud"* protocol applied it on three randomly selected articles.

Furthermore, before performing the actual inclusion and exclusion of studies, a pilot selection was performed (Petersen and Ali, 2011). This step was done by two reviewers independently on 20 randomly selected articles. The results of this pilot are shown in Table 3.

Table 3
Results of the pilot selection.

Reviewer 1	Reviewer 2		
	Relevant	Uncertain	Irrelevant
Relevant	6	0	2
Uncertain	–	4	0
Irrelevant	–	–	8

Table 4
Results of applying the inclusion and exclusion criteria.

Category ID	Number of articles	Total number of articles post discussion
A	96	106
B	34	34
C	122	174
D	30	0
E	82	0
F	1542	1592

We had an agreement on 90% of the 20 articles used in the pilot of inclusion/exclusion criteria. Based on these results with high level of agreement, we were confident to go ahead with the actual selection of the studies.

Inclusion and exclusion criteria were applied independently by two reviewers on 1906 articles that had passed the preliminary criteria used for initial screening (see Fig. 1).

The results of this phase are summarized in Table 4 where the third column shows the final total of articles once the articles in categories D and E were discussed and reclassified.

Table 5 shows good agreement on the outcome of applying basic criteria on articles. This shows a shared understanding and consistent application of the criteria on the articles. Only on 30 out of 1906 articles the reviewers had a major disagreement i.e. category D in Table 4.

Adaptive reading for articles in category C:

Based on the titles and abstracts of articles, we often lacked sufficient information to make a judgement about the context and method of the study. Therefore we had 174 articles (category C in Table 4) that required more information for decision making. Many of the existing literature reviews exclude such articles where both reviewers do not consider a study relevant (Petersen and Ali, 2011). However, as we decided to be more inclusive to minimize the threat of excluding relevant research we decided to further investigate such studies.

As the number of articles in this category was quite large (174 articles) and we already had a sizeable population of potential primary studies (106 and 34 articles in category A and B respectively) we could not justify spending a lot of effort in reading full-text of these articles. Therefore, we agreed on an appropriate level of

Table 5
Cohen's Kappa and percent agreement between reviewers.

Criteria	Percent agreement	Cohen's Kappa statistic
Basic criteria	92.50	0.73
Adaptive reading	78.60	0.53
Context description	80.50	0.65
Study design description	81.60	0.56
Validity threats discussion	95.40	0.73
Subjects/users	92.00	0.34
Scale	80.50	0.58
Model validity	89.70	0.83

detail to make a selection decision without having to read the full-text of the article. The resulting three-step process of inclusion and exclusion with increasing degree of detail is:

1. Read the introduction of the article to make a decision.
2. If a decision is not reached read the conclusion of the article.
3. If it is still unclear, search for the keywords and evaluate their usage to describe the context of the study in the article.

Again a pilot of this process was applied independently by the two reviewers on five randomly selected articles in category C. The reviewers logged their decisions and the step at which they took the decision e.g. *'Reviewer-1 has included article Y after reading its conclusion'*. In this pilot, we had a perfect agreement on four of the five articles with regard to the decision and the step where the decision was made. However, one article resulted in some discussion as reviewers noticed that in this article authors had used terms "empirical" and "example" and this made it unclear whether the study was done in real-world settings. To avoid exclusion of any relevant articles it was decided that such articles that are inconclusive in their use of these terms will be included for full-text reading.

The adaptive reading process described above was applied independently by both reviewers and we had a high congruence on what was considered relevant or irrelevant to the purpose of this study. The inter-rater agreement was fairly high for this step as presented in Table 5. All articles with conflicting decision between the two reviewers were taken to the next step for full-text reading. This resulted in another 74 articles for full-text reading in addition to the 140 articles in categories A and B see Table 4.

3.4.3. Advanced criteria

This is related to the actual data extraction, where the full-text of the articles was read by both reviewers independently. Exclude articles based on the same criteria used in the previous two steps (see Sections 3.4.1 and 3.4.2) but this time reading the full-text of the articles. We also excluded the conference articles that have been subsequently extended to journal articles (that are likely to have more details).

For practical reasons these 214 articles were divided equally among the two reviewers to be read in full-text. However, to minimize the threat of excluding a relevant study any article excluded by a reviewer was reviewed by the second reviewer. Section 3.7 presents the data extraction form used in this study, the results of the pilot and the actual data extraction performed in this study. The list of excluded studies at this stage are available in Ali et al. (2014).

3.5. Study quality assessment criteria

The criteria used in this study were adapted from Ivarsson and Gorschek (2010) to fit the area of SPSM. We dropped 'research methodology' and 'context' as criteria from the relevance category because we only included the real-world studies in this review. So, these fields were redundant.

3.5.1. Scoring for rigor

To assess how rigorously a study was done we used the following three sub-criteria:

- *Description of context*
  1. If the description covers at least four of the context facets: product; process; people; practices, tools, techniques; organization and market (Petersen and Wohlin, 2009) then the score is '1'.
  2. If the description covers at least two of the context facets then the score is '0.5'.
  3. If less than two facets are described then the score is '0'.

  In general, a facet was considered covered if even one of the elements related to a facet is described. The facet "process" was considered fulfilled if a general description of the process or if name of the process model followed in the organization is provided.
- *Study design description*
  1. If the data collection/analysis approach is described to be able to trace the following then the score is '1', which is given (a) what information source (roles/number of people/data set) was used to build the model, and (b) how the model was calibrated (variable to data-source mapping), and (c) how the model was evaluated (evaluation criteria and analysis approach).
  2. If data collection is only partially described (i.e. at least one of the three – (a), (b), or (c) above has been defined) then the score is '0.5'.
  3. If no data collection approach is described then the score is '0' (example: "we got the data from company X").
- *Discussion of validity threats*
  1. If all four types of threats to validity (Wohlin et al., 2012) (internal, external, conclusion and construct) are discussed then the score is '1'.
  2. If at least two threats of validity are discussed then the score is '0.5'.
  3. If less than two threats to validity are discussed then the score is '0'.

3.5.2. Scoring of relevance

The relevance of the studies for the software engineering practice was assessed by the following two sub-criteria: users/subjects and scale:

- *Users/subjects*
  1. If the intended users are defined and have made use of the simulation results for the purpose specified then the score is '1' (in case of prediction, e.g. a follow-up study or a post-mortem analysis of how it performed was done).
  2. If the intended users are defined and have reflected on the use of the simulation results for the purpose specified then the score is '0.5'.
  3. If the intended users have neither reflected nor made practical use of the model result then the score is '0' (e.g. the researcher just presented the result of the simulation and reflected on the output in the article).
- *Scale*
  1. If the simulation process is based on a real-world process then the score is '1' (articles that claim that the industrial process is similar to a standard process model were also scored as '1').
  2. If the simulation process has been defined by researchers without industry input then the score is '0' (the articles that only calibrate a standardized process model, will also get a zero).

To minimize the threat of researchers bias both reviewers performed the quality assessment of all the primary studies independently. Kappa statistic for inter-rater agreement was computed see Table 5. Generally we had a fair agreement as shown by the values of Cohen's Kappa (values greater than 0.21 are considered fair agreement). However, for criteria like subjects/users where we had a low agreement we do not think it is a threat to the validity of the results as all the conflicts were resolved by discussion and referring back to the full-text of the publication. The results of quality assessment of primary studies after consensus are given in Table A.1.

3.6. Scoring model validity

To assess the credibility of models (Ali and Petersen, 2012) developed and applied in the primary studies we used the following criteria:

1. If the following two steps were performed the model was scored as '1': (a) the model was presented to practitioners to check if it reflects their perception of how the process works (Kellner et al., 1999; Rus et al., 2003), or did sensitivity analysis (Kellner et al., 1999; Madachy, 2008); (b) checked the model against reference behaviour (Shull et al., 2007; Madachy, 2008) or compared model output with past data (Ahmed et al., 2005) or show model output to practitioners.
2. If at least one of (a) or (b) is reported then the score is '0.5'.
3. If there is no discussion of model verification and validation (V&V) then the score is '0'.

Both reviewers applied these criteria independently on all the primary studies. Cohen's Kappa value for inter-rater agreement for *"Model Validity"* is 0.83 (Table 5). This shows a high agreement between the reviewers and reliability of this assessment is also complemented by resolving all the disagreements by discussion and referring back to full-text of the publications.

3.7. Data extraction strategy

We used a random sample of 10 articles from the selected primary studies for piloting the data extraction form. The results were compared and discussed, this helped in developing a common interpretation of the fields in the data extraction form. This pilot also served to establish the usability of the form whether we did find the relevant information at all in the articles. The data extraction form had the following fields:

- **Meta information**: Study ID, author name, and title and year of publication.
- **Final decision**: Excluded if a study does not fulfil advanced criteria presented in Section 3.4.3.
- **Quality assessment**: Rigor (context description, study design description, validity discussion) and relevance (subjects/users, scale).
- **Model building**: Problem formulation (stakeholders, scope and purpose), simulation approach and tools used, data collection methods, model implementation, model verification and validation, model building cost, level of reuse, evaluation for usefulness, criteria and outcome of evaluation, documentation, and benefits and limitations.
- **Reviewer's own reflections**: The reviewers document notes, e.g. if an article has an interesting point that can be raised in the discussion.

We aimed to identify the intended purpose in the study as stated by their authors and not the potential/possible use of the simulation model in the study. In this regard, using the purpose statements extracted from the primary studies we followed the following three steps to aggregate the repeating purposes in the primary studies:

- **Step 1**: Starting with the first purpose statement create and log a code.
- **Step 2**: For each subsequent purpose statement identify if a purpose already exists. If it does log the statement with the existing code, otherwise create a new code.
- **Step 3**: Repeat Step 2 until the last statement has been catalogued.

The resulting clusters with same coded purpose were mapped to purpose categories defined in Kellner et al. (1999). However, we found that the purpose category "Understanding" overlaps with training and learning and it is so generic that it could be true for any simulation study no matter what was the purpose of the study.

Traceability was ensured between the mapping, clusters, and the purpose statements extracted from the primary studies. This enabled the second reviewer to review the results of the process above whether the statements were correctly clustered together. Any disagreements between the reviewers regarding the classification were resolved by discussion.

Similarly, the descriptive statements regarding the simulation model's scope that were extracted from the primary studies were analysed and mapped to the scopes identified by Kellner et al. (1999). This mapping was also reviewed by the second author for all the primary studies.

3.8. Validity threats

The threat of missing literature was reduced by using databases that cover computer science and software engineering. We further minimized the threat of not covering the population of relevant literature by doing a search in databases covering management related literature. Another step to ensure wide coverage was to consider all literature published before the year 2013 in this study. Thus, the search was not restricted by the time of publication or venue in any database.

Using the *"context"* block in the search string (as described in Section 3.3.2) adds a potential limitation to our search approach i.e. the articles that mention the name of the companies instead of the identified keywords, will not be found although they are industrial studies. However, this was a conscious decision as most often applied research is listed with the keywords used in this block. This was also alleviated to some an extent by using a broader search string in Google Scholar as described in Section 3.3.2. By using an electronic search (with search string) we reduced the selection bias (of reviewers) as well.

For practical reasons, the preliminary criteria were applied by one reviewer that may limit the credibility of the selection. However, by only removing the obviously outside the domain articles which were guided by simple and explicit criteria we tried to reduce this threat. Furthermore, at this stage and the later stages of selection we were always inclusive when faced with any level of uncertainty, this we consider also minimized the threat of excluding a relevant article. The selection of articles based on the basic criteria was done by both reviewers and an explicit strategy based on Petersen and Ali (2011) was employed.

All the selection of studies, data extraction procedures and quality assessment criteria were piloted and the results are presented in the paper. Any differences in pilots and actual execution of the studies were discussed and if needed the documentation of the criteria was updated based on the discussions. This was done to achieve consistency in the application of the criteria and to minimize the threat of misunderstanding by either of the reviewers. By making the criteria and procedure explicit, we have minimized the reviewer's bias and dependence of review results on personal judgements. This has further increased the repeatability of the review.

Inter-rater agreement was also calculated for such activities and is discussed in the paper where two reviewers performed a task e.g. application of basic criteria for selection and quality assessment. The inter-rater statistics reported in this study generally show a good agreement between reviewers. This shows that the criteria are explicit enough and support replication otherwise we would have had more disagreements. All the conflicts were resolved by discussion and reviewing the primary studies together. This means

that even on the criteria where the reviewers had a lower level of agreement it is not a threat to the results of the study. However, it does point out that the criteria were not explicit enough and is a threat to repeatability of the review.

Studies are in some cases based on Ph.D. theses, e.g. Abdel-Hamid (1988a,b). We evaluated the rigor and relevance based on what has been reported in the article, hence few studies that were based on the theses could potentially score higher. That is, the authors could have followed the step, but due to page restrictions did not report on the results. However, some of the studies only used calibration data and not the model itself. Given this situation, a few individual rigor and relevance scores could change, however, the principle conclusion would not be different.

Study selection and data analysis that resulted in classification of purpose and scope for the models also involved two reviewers to increase the reliability of the review. Explicit definition of criteria, and the experience of reviewers in empirical research in general and simulation in particular also increases the credibility of the review. Kitchenham et al. (2011) highlight the importance of research expertise in the area of review to increase the quality of study selection. The third author has used simulation in practice for software reliability and performance modelling for Telecommunication systems (Wohlin, 1991). First author has used simulation in industry practice to model the testing process and has a Licentiate on the topic (Ali and Petersen, 2012), and the second author has been part of the simulation projects performed by the first author.

4. Characteristics of studies

4.1. Number of new studies

Contrary to earlier research we identified significantly more studies from the real-world software development context. Zhang et al. (2011) stated that they found *"32 industrial application cases"* of which *"given the limited space, this paper, as an initial report, only describes some of the important SPS application cases we identified"*. Similarly in a systematic literature review of software process modelling literature that included both static and dynamic modelling they found a combined total of only 43 articles (Bai et al., 2011).

4.2. Purpose

Table 6 gives an overview of real-world simulation studies relating to the purposes defined in Kellner et al. (1999). The clear majority of studies used simulation for planning purposes (45 studies), followed by process improvement (26 studies), and training and learning (21 studies). In comparison, only a few studies used simulation for control and operational management.

Planning: In planning simulation has been used for decision support (Abdel-Hamid, 1988a, 1989a, 1993; Shen et al., 2005; Huang et al., 2006; Ghosh and Wei, 2010; Dickmann et al., 2007; Regnell et al., 2004; Rus et al., 2003; Meilong et al., 2008) (e.g. in relation to staffing and allocation of effort (Rus et al., 2003; Abdel-Hamid, 1989a; Ghosh and Wei, 2010), connecting decisions and simulating them in relation to business outcomes (Dickmann et al., 2007), and requirements selection (Regnell et al., 2004)). Furthermore, simulation has been used by many studies for estimation (Antoniades et al., 2002; Park et al., 2008b; Abdel-Hamid, 1990; McCall et al., 1979; Bai et al., 2009, 2009; Mizuno et al., 2001, 1997; Mizell and Malone, 2007a,b; Pfahl and Lebsanft, 2000b; Madachy and Khoshnevis, 1997; Lin and Levary, 1989; Tausworthe and Lyu, 1994; Höst et al., 2001; Wiegner and Nof, 1993; Setamanit and Raffo, 2008; Al-Emran et al., 2008), some examples are cost and effort estimation (McCall et al., 1979; Bai et al., 2009; Mizell and Malone, 2007a), schedule estimation (Abdel-Hamid, 1990), release planning, and fault/quality estimation (Tausworthe and Lyu, 1994; Madachy and Khoshnevis, 1997; Kusumoto et al., 2001).

Table 6
Purpose of simulation in the primary studies.

Purpose	Number of articles	References
Control and operational management	9	McCall et al. (1979), Hsueh et al. (2008), Antoniol et al. (2004), Yong and Zhou (2010), Wu et al. (2010), Pfahl and Lebsanft (2000a), Zhang et al. (2008e), Raffo (2005) and Paikari et al. (2012)
Planning	45	Abdel-Hamid (1988a, 1989a,b, 1990, 1993), Abdel-Hamid and Leidy (1991), Madachy and Khoshnevis (1997), Mizell and Malone (2007a,b), Kusumoto et al. (2001), Dickmann et al. (2007), Antoniades et al. (2002), Shen et al. (2005), Tausworthe and Lyu (1994), Wang and Chen (2003), McCall et al. (1979), Huang et al. (2006), Lin and Levary (1989), Ghosh and Wei (2010), Rus et al. (2003), Park et al. (2008b), Höst et al. (2001), Mizuno et al. (1997, 2001), Ferreira et al. (2009), Pfahl and Lebsanft (2000a,b), Houston et al. (2001), Meilong et al. (2008), Wiegner and Nof (1993), Al-Emran et al. (2008, 2010), Madachy (1995), Regnell et al. (2004), Setamanit and Raffo (2008), Bai et al. (2009), Lin (2011), Ba and Wu (2012), Dasgupta et al. (2011), Spasic and Onggo (2012), Junjie et al. (2012), Jian-Hong et al. (2011), Di Penta et al. (2011), Farshchi et al. (2012) and Crespo and Ruiz (2012)
Process improvement and technology adoption	26	Car et al. (2002, 2010), Pfahl et al. (2004), Raffo et al. (1999, 2004, 2008), Houston (2006), Martin and Raffo (2001), Melis et al. (2006), Ferreira et al. (2009), Ruiz et al. (2004), Aranha and Borba (2008), Nakatani and Nishida (1992), Lin et al. (1997), Wernick and Lehman (1999), Stallinger (2000), Chiang and Menzies (2002), Podnar and Mikac (2001), Turnu et al. (2006), Li et al. (2006), Deissenboeck and Pizka (2008), Rahmandad and Weiss (2009), Anderson et al. (2012), Seunghun and Doo-Hwan (2011), Gillenson et al. (2011) and Psaroudakis and Eberhardt (2011)
Training and learning	21	Abdel-Hamid (1988b), Ruiz et al. (2001, 2004), Birkhölzer et al. (2005), Hsueh et al. (2008), Park et al. (2008b), Höst et al. (2001, 2001), Andersson et al. (2002), Ferreira et al. (2009), Katsamakas and Georgantzas (2007), Ramil and Smith (2002), Christie and Staley (2000), Chen and Wei (2009), Chatters et al. (2000), Setamanit and Raffo (2008), Bai et al. (2009), Zhang (2009), Rahmandad and Weiss (2009), Anderson et al. (2012) and Car et al. (2010)

Process improvement and technology adoption: Studies in this category used simulation to evaluate alternative process designs for process improvements (Martin and Raffo, 2001; Podnar and Mikac, 2001; Stallinger, 2000; Pfahl et al., 2004; Raffo et al., 1999; Wernick and Lehman, 1999; Houston, 2006; Lin et al., 1997; Li et al., 2006). As an example, (Stallinger, 2000) investigated the effect of conducting an improvement plan driven by the ISO/IEC 15504 standard. Furthermore, improvements have been evaluated by varying a number of parameters in the process to determine the best alternatives one ought to strive for (cf. Wernick and Lehman, 1999; Houston, 2006; Lin et al., 1997), as well as investigating specific technology adoptions to the process (Raffo et al., 2008, 2004; Deissenboeck and Pizka, 2008; Melis et al., 2006; Turnu et al., 2006; Aranha and

Borba, 2008; Chiang and Menzies, 2002). Examples of technology adoptions were introduction of test driven development (TDD) (Melis et al., 2006; Turnu et al., 2006), comparison of manual vs. model-based techniques (Aranha and Borba, 2008), or use of different quality assurance approaches or architectural styles (Chiang and Menzies, 2002).

Training and learning: In training and learning the majority of the studies aimed to explore or understand a phenomena from a scientific point of view to provide some recommendations and guidelines to practitioners (cf. Ruiz et al., 2001, 2004; Höst et al., 2001, 2001; Ferreira et al., 2009; Park et al., 2008b; Zhang, 2009; Setamanit and Raffo, 2008; Katsamakas and Georgantzas, 2007; Abdel-Hamid, 1988b; Chatters et al., 2000; Bai et al., 2009). Examples are to understand the effect of requirements overload on bottlenecks in the whole process (Höst et al., 2001) or understanding open source system development (Katsamakas and Georgantzas, 2007), assess what factors make global software development successful (Setamanit and Raffo, 2008), or understanding the effects of creeping requirements (Park et al., 2008b).

Control and operational management: Only few studies used simulation for control and operational management (Hsueh et al., 2008; Pfahl and Lebsanft, 2000a; Antoniol et al., 2004; Wu et al., 2010; Zhang et al., 2008e; McCall et al., 1979; Yong and Zhou, 2010). As an example, several studies assessed whether a project is likely to meet its expected deadline (Antoniol et al., 2004; Wu et al., 2010), or whether an individual iteration is able to meet the deadline (Zhang et al., 2008e). Studies (McCall et al., 1979; Yong and Zhou, 2010) used simulation for progress status assessment.

4.3. Scope

Table 7 defines categories for scope that a simulation study can have. The clear majority of studies focused on individual projects (56 studies). In comparison, fewer studies looked at the portion of a lifecycle (21 studies), such as testing. Only a small subset of studies investigated simulations in the context of long term evolution (4 studies), long term organization (3 studies), and concurrent projects (3 studies).

Projects: Studies investigating projects and their development lifecycle looked at different lifecycles, e.g. related to CMMI processes (Hsueh et al., 2008), and extreme programming (XP) projects (Melis et al., 2006). One study explicitly stated that they are focusing on new software development (Rus et al., 2003), while others in connection with the software lifecycle (excluding requirements and maintenance phase, see e.g. Abdel-Hamid, 1988b, 1990). The remaining studies only specified that they are looking at the overall development lifecycle of projects without further specification of the type and scope/boundaries.

A portion of the lifecycle: Studies looking at a portion of a lifecycle investigated focused on the maintenance process solely (Car et al., 2002; Podnar and Mikac, 2001; Ramil and Smith, 2002), software reliability lifecycle (Tausworthe and Lyu, 1994), requirements and test (Andersson et al., 2002), quality assurance processes (Aranha and Borba, 2008; Raffo et al., 2004), requirements (Christie and Staley, 2000), release processes (Höst et al., 2001), and processes for components off the shelf (COTS) selection and use (Ruiz et al., 2004).

Long term product evolution: Researchers have looked at the general long-term evolution of products without more specific classification (Wernick and Lehman, 1999; Chatters et al., 2000; Zhang, 2009) while (Regnell et al., 2004) looked at market-driven requirements processes from a long-term perspective.

Long term organization: From an organizational perspective, studies investigated factors influencing process leadership (Pfahl et al., 2004), processes at an organizational level (Dickmann et al., 2007), and CMMI process in relation to organizational business concerns (Birkhölzer et al., 2005).

Table 7
Scope of simulation models in the primary studies.

Scope	Number of articles	References
A portion of lifecycle	21	Car et al. (2002), Tausworthe and Lyu (1994), Houston (2006), Antoniol et al. (2004), Andersson et al. (2002), Anderson et al. (2012), Ruiz et al. (2004), Aranha and Borba (2008), Raffo et al. (2004), Podnar and Mikac (2001), Ramil and Smith (2002), Al-Emran et al. (2008, 2010), Christie and Staley (2000), Höst et al. (2001), Lin (2011), Ba and Wu (2012), Seunghun and Doo-Hwan (2011), Spasic and Onggo (2012), Paikari et al. (2012) and Di Penta et al. (2011)
Development project	56	Abdel-Hamid (1988a,b, 1989a,b, 1990), Abdel-Hamid and Leidy (1991), Madachy and Khoshnevis (1997), Mizell and Malone (2007a,b), Kusumoto et al. (2001), Raffo et al. (1999, 2008), Raffo (2005), Antoniades et al. (2002), Shen et al. (2005), Ruiz et al. (2001), Wang and Chen (2003), McCall et al. (1979), Martin and Raffo (2001), Huang et al. (2006), Hsueh et al. (2008), Lin and Levary (1989), Ghosh and Wei (2010), Rus et al. (2003), Park et al. (2008b), Melis et al. (2006), Mizuno et al. (1997, 2001), Ferreira et al. (2009), Pfahl and Lebsanft (2000a,b), Katsamakas and Georgantzas (2007), Nakatani and Nishida (1992), Lin et al. (1997), Houston et al. (2001), Meilong et al. (2008), Wiegner and Nof (1993), Chiang and Menzies (2002), Yong and Zhou (2010), Wu et al. (2010), Turnu et al. (2006), Chen and Wei (2009), Li et al. (2006), Madachy (1995), Zhang et al. (2008e), Setamanit and Raffo (2008), Deissenboeck and Pizka (2008), Bai et al. (2009), Dasgupta et al. (2011), Junjie et al. (2012), Gillenson et al. (2011), Car et al. (2010), Jian-Hong et al. (2011), Farshchi et al. (2012), Psaroudakis and Eberhardt (2011) and Crespo and Ruiz (2012)
Concurrent projects	3	Abdel-Hamid (1993), Stallinger (2000) and Rahmandad and Weiss (2009)
Long term evolution	4	Wernick and Lehman (1999), Chatters et al. (2000), Regnell et al. (2004) and Zhang (2009)
Long term organization	3	Pfahl et al. (2004), Dickmann et al. (2007) and Birkhölzer et al. (2005)

Concurrent projects: Two parallel projects have been simulated by Abdel-Hamid (1993), while multiple concurrent projects have been simulated by Stallinger (2000) and Rahmandad and Weiss (2009).

4.4. Simulation approaches

Table 8 shows the simulation approaches used by the identified studies. System dynamics (SD) is the most commonly used simulation approach (32 studies), followed by discrete event simulation (DES) with 22 studies. A total of 12 studies combined different simulation approaches. Only few studies used approaches such as Petri nets (PN) (6 studies) and Monte Carlo (3 studies).

Hybrid simulation models were mostly combining SD and DES (Martin and Raffo, 2001; Park et al., 2008b; Melis et al., 2006;

Table 8
Simulation approaches used in the primary studies.

Approach	Number of articles	References
DES	23	Car et al. (2002), Mizell and Malone (2007a,b), Tausworthe and Lyu (1994), Houston (2006), Wang and Chen (2003), Birkhölzer et al. (2005), Antoniol et al. (2004), Raffo et al. (1999, 2004, 2008), Podnar and Mikac (2001), Al-Emran et al. (2008), Höst et al. (2001), Regnell et al. (2004), Nakatani and Nishida (1992), Lin (2011), Raffo (2005), Seunghun and Doo-Hwan (2011), Junjie et al. (2012), Di Penta et al. (2011), Psaroudakis and Eberhardt (2011) and Dickmann et al. (2007)
Hybrid	12	Martin and Raffo (2001), Park et al. (2008b), Melis et al. (2006), Ferreira et al. (2009), Ramil and Smith (2002), Christie and Staley (2000), Zhang et al. (2008e), Setamanit and Raffo (2008), Bai et al. (2009), Zhang (2009), Crespo and Ruiz (2012) and Al-Emran et al. (2010)
Monte Carlo	3	Ghosh and Wei (2010), Chiang and Menzies (2002) and Dasgupta et al. (2011)
Other	11	Wiegner and Nof (1993), Wu et al. (2010), Deissenboeck and Pizka (2008), Chen and Wei (2009), McCall et al. (1979), Aranha and Borba (2008), Hsueh et al. (2008), Spasic and Onggo (2012), Anderson et al. (2012), Jian-Hong et al. (2011) and Gillenson et al. (2011)
PN	6	Kusumoto et al. (2001), Shen et al. (2005), Huang et al. (2006), Mizuno et al. (1997, 2001) and Li et al. (2006)
SD	32	Abdel-Hamid (1988a,b, 1989a,b, 1990, 1993), Abdel-Hamid and Leidy (1991), Madachy and Khoshnevis (1997), Pfahl and Lebsanft (2000a,b), Pfahl et al. (2004), Antoniades et al. (2002), Ruiz et al. (2001, 2004), Lin and Levary (1989), Rus et al. (2003), Andersson et al. (2002), Katsamakas and Georgantzas (2007), Lin et al. (1997), Houston et al. (2001), Meilong et al. (2008), Wernick and Lehman (1999), Stallinger (2000), Yong and Zhou (2010), Turnu et al. (2006), Madachy (1995), Chatters et al. (2000), Ba and Wu (2012), Rahmandad and Weiss (2009), Car et al. (2010), Paikari et al. (2012) and Farshchi et al. (2012)

Christie and Staley, 2000; Zhang et al., 2008e; Setamanit and Raffo, 2008; Bai et al., 2009). Furthermore, qualitative simulation was combined with other models such as SD, or used to abstract from a quantitative model (cf. (Zhang, 2009; Ramil and Smith, 2002)). Furthermore, models combined stochastic and deterministic aspects in a single model (Ferreira et al., 2009).

Others include a variety of models that used approaches from control theory (Wiegner and Nof, 1993), Copula methods (Wu et al., 2010), Markov Chaims (Deissenboeck and Pizka, 2008), agent-based (Chen and Wei, 2009; Spasic and Onggo, 2012; Anderson et al., 2012), while others did not specify the approach used, and we could not deduce the approach from the information presented.

In connection to simulation approaches, it is interesting which tools have been used to implement them.

SD: Three studies reported the use of Vensim (Ruiz et al., 2004; Pfahl and Lebsanft, 2000a,b). Other simulation tools used were Powersim (Andersson et al., 2002), and iThink (Madachy and Khoshnevis, 1997). One study used a self-developed tool (Madachy, 1995). The majority of SD studies did not report the tool they used.

DES: For DES two studies used Extend (Raffo et al., 2004; Al-Emran et al., 2008) and two studies used Process Analysis Tradeoff Tool (PATT) (Mizell and Malone, 2007a,b). Further tools used were Micro Saint (Car et al., 2002), RSQsim (Regnell et al., 2004), Telelogic SDL modelling tool (Höst et al., 2001) and SoftRel (Tausworthe and Lyu, 1994). For two studies simulation models were programmed from scratch in Java (Birkhölzer et al., 2005) and C++ (Antoniol et al., 2004). One study used Statemate Magnum (Raffo et al., 1999) and Nakatani (Nakatani and Nishida, 1992) did not specify the tool used.

Hybrid: For hybrid simulation a variety of tools have been used, such as DEVSim++ (APMS) (Park et al., 2008b), Extend (Christie and Staley, 2000), iThink (Ferreira et al., 2009), Little-JIT (Bai et al., 2009), QSIM (Ramil and Smith, 2002), and Vensim (Zhang, 2009). One study used a self-developed model written in Smalltalk (Melis et al., 2006).

PN: One study documented that they developed the model from scratch using C (Kusumoto et al., 2001), others did not report on the tools used.

Monte Carlo: No information about tools has been provided.

4.5. Cross analysis purpose and scope

Table 9 shows a cross-analysis of purpose and scope. The simulation of individual development projects is well covered across all purposes with multiple studies for each purposes.

A portion of the lifecycle was primarily investigated for process improvement and technology adoption (8 studies) and planning (8 studies). A few studies investigated training and learning (6 studies), and only two studies focused on control and operational management.

Overall, research on concurrent projects is scarce. One study investigated concurrent projects for planning, two for process improvement and technology adoption, and one for training and learning.

Similar patterns are found for long term evolution and long term organization, where primarily individual studies investigated the different purposes with respect to the defined scopes of the simulation models.

No studies in the area of control and operational management focus on long term evolution or organization, which is by definition to be expected.

4.6. Cross analysis purpose and simulation approaches

Table 10 shows a cross-analysis of purpose and simulation approaches. SD has been used for all purposes, while an emphasis is given to planning (17 studies), followed by process improvement (8 studies) and training and learning (8 studies). Only three SD studies focused on control and operational management.

Hybrid simulation has been used for all purposes as well, with the majority of studies focusing on training and learning (7 studies), followed by planning (6 studies), and process improvement (3 studies). Only one hybrid study has been used for control and operational management.

Overall, the number of Monte Carlo simulations has been low with only three studies. Out of these three, two studies focus on planning and one on process improvement.

PN studies primarily used the simulation for planning purposes, only one study focused on process improvement.

Others have been used in a balanced way across purposes.

Table 9
Purpose and scope of SPSM in the primary studies.

	Control and operational management	Planning	Process improvement and technology adoption	Training and learning
A portion of lifecycle	Antoniol et al. (2004) and Paikari et al. (2012)	Tausworthe and Lyu (1994), Höst et al. (2001), Al-Emran et al. (2010), Ba and Wu (2012), Di Penta et al. (2011), Lin (2011) and Spasic and Onggo (2012)	Car et al. (2002), Houston (2006), Ruiz et al. (2004), Aranha and Borba (2008), Raffo et al. (2004), Podnar and Mikac (2001), Anderson et al. (2012) and Seunghun and Doo-Hwan (2011)	Andersson et al. (2002), Ruiz et al. (2004), Ramil and Smith (2002), Christie and Staley (2000), Höst et al. (2001) and Anderson et al. (2012)
Development project	McCall et al. (1979), Hsueh et al. (2008), Yong and Zhou (2010), Wu et al. (2010), Pfahl and Lebsanft (2000a), Zhang et al. (2008e) and Raffo (2005)	Abdel-Hamid (1988a, 1989a,b, 1990), Abdel-Hamid and Leidy (1991), Madachy and Khoshnevis (1997), Mizell and Malone (2007a,b), Kusumoto et al. (2001), Antoniades et al. (2002), Shen et al. (2005), Wang and Chen (2003), McCall et al. (1979), Huang et al. (2006), Lin and Levary (1989), Ghosh and Wei (2010), Rus et al. (2003), Park et al. (2008b), Mizuno et al. (1997, 2001), Ferreira et al. (2009), Pfahl and Lebsanft (2000a,b), Houston et al. (2001), Meilong et al. (2008), Wiegner and Nof (1993), Madachy (1995), Setamanit and Raffo (2008), Bai et al. (2009), Crespo and Ruiz (2012), Dasgupta et al. (2011), Farshchi et al. (2012), Jian-Hong et al. (2011), Junjie et al. (2012) and Al-Emran et al. (2008)	Raffo et al. (1999, 2008), Martin and Raffo (2001), Melis et al. (2006), Ferreira et al. (2009), Nakatani and Nishida (1992), Lin et al. (1997), Chiang and Menzies (2002), Turnu et al. (2006), Li et al. (2006), Deissenboeck and Pizka (2008), Car et al. (2010), Gillenson et al. (2011) and Psaroudakis and Eberhardt (2011)	Abdel-Hamid (1988b), Ruiz et al. (2001), Hsueh et al. (2008), Park et al. (2008b), Ferreira et al. (2009), Katsamakas and Georgantzas (2007), Chen and Wei (2009), Setamanit and Raffo (2008), Bai et al. (2009) and Car et al. (2010)
Concurrent projects		Abdel-Hamid (1993)	Stallinger (2000) and Rahmandad and Weiss (2009)	Rahmandad and Weiss (2009)
Long term evolution		Regnell et al. (2004)	Wernick and Lehman (1999)	Chatters et al. (2000) and Zhang (2009)
Long term organization		Dickmann et al. (2007)	Pfahl et al. (2004)	Birkhölzer et al. (2005)

5. Systematic literature review results

To assess the evidence of usefulness of simulation for the proposed purpose we would like to take into account the rigor and relevance of studies, model's credibility (in terms of the level of verification and validation), scope of the model and the real-world context in which it was used.

5.1. Context of simulation studies

Petersen and Wohlin (2009) defined different facets to describe the industrial context of a study. Each facet contains elements that could be described to characterize the facet, serving as a checklist for researchers during reporting of studies. For this study, a characterization of the following facets was sought: (1) product, (2) processes, (3) people, (4) practices, tools, and techniques, (5) product, (6) organization, and (7) market. Among the primary studies only 9% (8 studies) cover at least four context facets (Petersen and Wohlin, 2009) and 56% have described less than two context facets in the articles.

5.2. Verification and validation of simulation models

To validate the model structure and representativeness:

- 16% of the studies used *"practitioner feedback"*.
- Only six studies i.e. 7% reported performing sensitivity analysis on the model.
- While 76% studies did not report any such effort.

It can be said that in some of these studies this level of validation of representativeness may not have been necessary since the models were based on standards e.g. the IEEE-12207 (under the premise that the organization uses a similar process).

To validate the model behaviour:

- 47% of the studies compared the simulation model output with real data.
- 5% (i.e. four studies) reviewed the model output with practitioners.
- 6% (i.e. five studies) either compared the model output with literature or with other models.
- 40% reported no such effort.

5.3. Rigor and relevance of primary studies

Fig. 2 provides an overview of how the primary studies in this review scored against the rigor-relevance criteria. From this figure, we can divide the studies into four clear sets as listed below. For example, studies are classified as 'A' (high rigor, low relevance) if they have a rigor score above the median value '1.5' and a relevance score below or equal to the median value '1'.

Table 10
Purpose and approach of SPSM in the primary studies.

	Control and operational management	Planning	Process improvement and technology adoption	Training and learning
DES	Antoniol et al. (2004) and Raffo (2005)	Mizell and Malone (2007a,b), Tausworthe and Lyu (1994), Wang and Chen (2003), Al-Emran et al. (2008), Höst et al. (2001), Regnell et al. (2004), Di Penta et al. (2011), Junjie et al. (2012), Lin (2011) and Dickmann et al. (2007)	Car et al. (2002), Houston (2006), Raffo et al. (1999, 2004, 2008), Podnar and Mikac (2001), Psaroudakis and Eberhardt (2011), Seunghun and Doo-Hwan (2011) and Nakatani and Nishida (1992)	Birkhölzer et al. (2005) and Höst et al. (2001)
Hybrid	Zhang et al. (2008e)	Park et al. (2008b), Ferreira et al. (2009), Setamanit and Raffo (2008), Bai et al. (2009), Crespo and Ruiz (2012) and Al-Emran et al. (2010)	Martin and Raffo (2001), Melis et al. (2006) and Ferreira et al. (2009)	Park et al. (2008b), Ferreira et al. (2009), Ramil and Smith (2002), Christie and Staley (2000), Setamanit and Raffo (2008), Bai et al. (2009) and Zhang (2009)
Monte Carlo		Ghosh and Wei (2010) and Dasgupta et al. (2011)	Chiang and Menzies (2002)	
Other	McCall et al. (1979), Hsueh et al. (2008) and Wu et al. (2010)	McCall et al. (1979), Wiegner and Nof (1993), Spasic and Onggo (2012) and Jian-Hong et al. (2011)	Aranha and Borba (2008), Deissenboeck and Pizka (2008), Anderson et al. (2012) and Gillenson et al. (2011)	Hsueh et al. (2008), Chen and Wei (2009) and Anderson et al. (2012)
PN		Kusumoto et al. (2001), Shen et al. (2005), Huang et al. (2006) and Mizuno et al. (1997, 2001)	Li et al. (2006)	
SD	Yong and Zhou (2010), Pfahl and Lebsanft (2000a) and Paikari et al. (2012)	Abdel-Hamid (1988a, 1989a,b, 1990, 1993), Abdel-Hamid and Leidy (1991), Madachy and Khoshnevis (1997), Antoniades et al. (2002), Lin and Levary (1989), Rus et al. (2003), Pfahl and Lebsanft (2000a,b), Houston et al. (2001), Meilong et al. (2008), Madachy (1995), Ba and Wu (2012) and Farshchi et al. (2012)	Pfahl et al. (2004), Ruiz et al. (2004), Lin et al. (1997), Wernick and Lehman (1999), Stallinger (2000), Turnu et al. (2006), Car et al. (2010) and Rahmandad and Weiss (2009)	Abdel-Hamid (1988b), Ruiz et al. (2001), Andersson et al. (2002), Ruiz et al. (2004), Katsamakas and Georgantzas (2007), Chatters et al. (2000), Car et al. (2010) and Rahmandad and Weiss (2009)

- 81 studies classified as 'C': with (low rigor, low relevance) of ($\leq$1.5, $\leq$1).
- 4 studies classified as 'B': with (low rigor, high relevance) of ($\leq$1.5, >1).
- 2 studies classified as 'A': with (high rigor, low relevance) of (>1.5, $\leq$1).
- 0 studies classified as 'D': with (high rigor, high relevance) of (>1.5, >1).

Furthermore, Fig. 2 shows the median age of the articles in each category. It is visible that the quality of the studies does not seem to exhibit a trend of increase in rigor and relevance scores in relation to newer articles.

5.4. Evaluation of simulation for the intended purpose

Only 13% (11 studies) actually reported some sort of evaluation of the model's usefulness for the suggested purpose. 6% of studies compared predicted data to real-data (which we have considered an acceptable evaluation provided it is the prediction and not just replication of the real-data based on the calibration of the model). Of the studies 6% used feedback from practitioners for the proposed model purposes (however not based on actual use of the model). Only one study (Houston, 2006) reports a follow-up study where the effectiveness of simulation for the purpose of software process improvement is investigated.

6. Discussions

6.1. Coverage of modelling scope and purposes

Overall, studies taking a long-term perspective are under-represented, only four studies looked at long term evolution, and three covered long term organization (see Table 9). Furthermore, there are only three studies reporting on simulating concurrent projects, while these become more relevant at this point in time. In particular, in system of systems development and large-scale software development taking a perspective beyond the project is important. From a research perspective, this implies

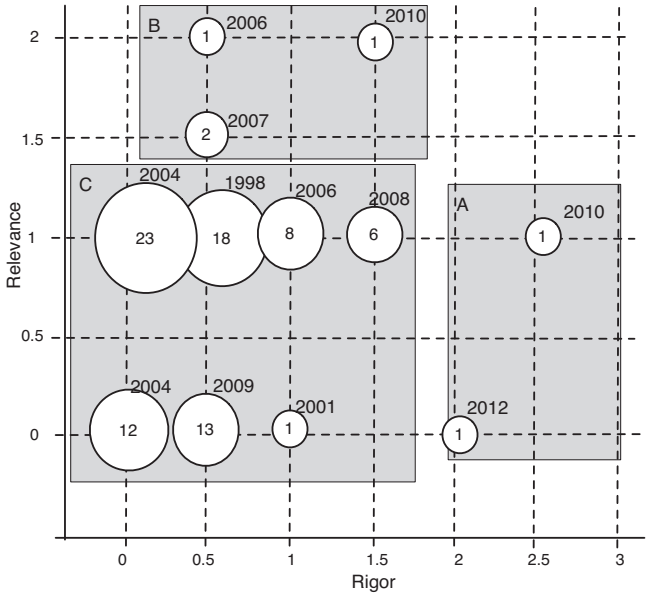

Fig. 2. Overview of rigor-relevance scores for the primary studies.

the need for applying simulation on concurrent projects from an end to end development lifecycle. Looking at the intersection between purpose and lifecycle, it becomes more apparent that long-term perspective, long term organization, and concurrent project simulation is a research gap. Only individual studies in those areas focused on planning and process improvement/technology adoption. A cross-analysis of purpose and simulation approaches revealed that overall the different simulation approaches were applied across all the purposes.

6.2. Reporting quality

23% of the primary studies partially describe the data collection, analysis and evaluation approach. 62% of these studies do not report the study design in appropriate detail (the criteria was discussed in Section 3.5). Only two studies explained the study design in detail. Among the four typical threats to validity (Wohlin et al., 2012) in empirical research only 8% of the primary studies had a discussion of two or three threats to validity. Remaining 91% of the primary studies did not have any discussion on validity threats.

This gives a clear picture of the quality of reporting in primary studies and makes it difficult to analyse the credibility and the strength of evidence in these studies. From a secondary studies perspective, this highlights the challenge to synthesize the evidence (if any) reported in these studies.

6.3. Context

It is critical to identify the contextual factors because their interaction influences what works well in a certain setting. It is crucial to contextualize empirical evidence by aptly raising the question *"What works for whom, where, when and why?"* (Petersen and Wohlin, 2009; Dybå, 2013). Unfortunately, given that 56% of the articles describe less than two facets of context it is impossible to perform an analysis where we may establish a relation between the usefulness of simulation approaches in a given context using the empirical evidence from the primary studies.

6.4. Model validity

Given the lack of empirical data in industrial contexts one would tend to agree with Dickmann et al. (2007) when they state that, *"Methodologically, the simplest case is to prove congruence of the simulation results with real world input-output-data. However, this type of validation is almost never achieved because of "lack of data" from software development projects"*. However, surprisingly 51% of the primary studies reported a comparison of the simulation model output with real data for validation.

One reason could be that the SPSM study's goals were aligned with the available data. Another explanation could be that the companies where the studies were conducted have extensive measurement programs. The first explanation is highly likely for studies where a phase of the development process is modelled with a narrow focus e.g. the simulation of maintenance process (Car et al., 2002). The second may also hold true as 21% of the studies either simulated a process at NASA or used their data to calibrate the model (Abdel-Hamid, 1988a,b, 1989a, 1990, 1993; Abdel-Hamid and Leidy, 1991; Mizell and Malone, 2007a,b; Raffo et al., 2008, 2004; Lin et al., 1997; Chiang and Menzies, 2002; Christie and Staley, 2000) since they had a strong tradition of measurements.

Overall in terms of V&V of the simulation models built and used in the primary studies, 26% had not reported any V&V whereas 45% had some level of V&V and 16% had reported V&V of both the model's structure and behaviour. From the simulation literature (Madachy, 2008; Law, 2001) we know that the credibility of simulation models cannot be guaranteed by just one of the methods of V&V. This also points to poor credibility of the evidence reported in simulation studies. As pointed out by Ghosh and Wei (2010) reproducing the output correctly does not validate the underlying cause effect relations in the model. Out of the secondary studies (discussed in Section 2.1), only França and Travassos (2013) have critiqued existing research for lacking rigor in model validation and the use of analysis procedures on simulation output data. Our study independently concurs their findings with respect to the model validation (França and Travassos, 2013).

6.5. Evaluation of usefulness

Of the 87 primary studies we found a diverse coverage of simulation purpose, model scope and simulation approaches. This was summarized in Sections 4.5 and 4.6. However, from the literature review's perspective, the lack of evaluation of simulation models for proposed purposes is disappointing.

Ideally speaking we would like to use the evidence in studies that score high on both rigor and relevance. However, the primary studies in this review scored poorly on both dimensions. Let us therefore look at the studies in categories 'A' and 'B' (that scored high on at least one dimension as reported in Section 3.5) that also performed evaluation of simulation models for their intended purposes. We have three such 'B' studies and one 'C' study.

Class 'A':

The two studies (Ghosh and Wei, 2010; Junjie et al., 2012) though with high rigor yet low relevance scores did not perform an evaluation of the model for the proposed purpose at all.

Class 'B':

Hsueh et al. (2008) conducted a questionnaire based survey to assess if the model can be useful for educational purposes. Although they report that most subjects expressed that the model *"could provide interesting and effective process education scenarios"*, but 25% of respondents considered the simulation based games were not helpful or useful to their current work. This can only be considered as an initial evaluation of the approach and given the overall positive outcome a thorough evaluation of the game for educational purpose is required.

Huang et al. (2006) report that the project managers claimed that they would have made a value-neutral decision without the (simulation based) analysis from the study. However, in our opinion it is not clear if the contribution is that of the simulation model or that of the underlying *"Value Based Software Quality Achievement"* process framework.

Houston (2006) uses DES to quantitatively analyse process improvement alternatives and then performs a follow-up study to see if the changes done in the process resulted in real improvements. Although there is no discussion on how the confounding factors and threats to the validity of the study were mitigated it is difficult to associate the evidence of improvements in the process with use of process simulation to assess improvement alternatives.

Al-Emran et al. (2010) combines Monte Carlo simulation with DES and models the release planning phase. Without presenting any details of the evaluation, they claimed that it was found useful for early risk detection and the development of mitigation strategies in the case study conducted at a company. Furthermore, validation of neither the structure nor the behaviour of the model is reported in the article.

The aspect of not performing evaluations may have permeated into SPSM from other disciplines that use simulation, where the emphasis is on verification and validation to ensure that the results are credible. In such disciplines, the results are often considered sufficiently credible if they are accepted and implemented by the sponsor of a simulation study (Balci, 1990). However, it should be considered that these disciplines have an established tradition of using simulation and ample success-cases exist for them not to do follow-up studies.

Whereas to establish SPSM as an alternative to static process models, analytical models, and as means of planning, management and training in industry, the evidence of its efficacy must be presented. For example in the context of lean development, comparing a value stream mapping workshop (Mujtaba et al., 2010) with static process descriptions with one facilitated by a dynamic simulation model. Or evaluating the effectiveness of simulation based training of practitioners compared to a well conducted seminar with graphical aids.

Unfortunately, based on the results of this systematic review and the modellers survey (Ahmed et al., 2008), it is indicated that SPSM modellers see verification and validation as the evaluation of simulation models. Some initial work towards a framework that highlights the need for evaluation of the value aspect of a simulation model is done by Ahmed et al. (2003).

6.6. Comparison with previous literature reviews

The total number of simulation studies (87 identified in this study) that were related to real-world application of SPSM is fairly high, which is not indicated by previous literature reviews on simulation. For example, Zhang et al. (2011) stated that they found *"32 industrial application cases"* of which *"given the limited space, this paper, as an initial report, only describes some of the important SPS application cases we identified"*. França and Travassos (2013) reported on the frequency of studies with respect to domain, which partially overlaps with what we defined as the scope. A shared observation is that the most frequent investigations relate to development projects. Furthermore, de França and Travassos reported on verification and validation, in particular of 108 of their included studies 17 papers compared their results with actual results, 14 had model results reviewed by experts, and 3 studies used surveys to confirm the validity of model behaviour. Whether studies do combinations of them cannot be deduced from the tables presented. However, the study confirms that only a small portion of studies conducts verification and validation of models. A noteworthy difference between França and Travassos (2013) and the study presented in this paper is the difference in population and sampling. Their population focused on all types of simulation in software engineering (including architecture simulation), and in their sampling they did not include business literature. Lastly Bai et al. (2011) have identified a total of 43 empirical studies even though their population was all software process modelling literature including SPSM.

Furthermore, none of the existing systematic reviews (Liu et al., 2009; Zhang et al., 2008a,b,d, 2010; Bai et al., 2011) report the lack of evidence for the usefulness of SPSM in current research. Zhang et al. (2011) do indicate a lack of objective evaluation but do not provide any traceable foundation for their claim and they do not highlight the almost non-existence of evidence for the usefulness of SPSM.

6.7. Cost of SPSM

When proposing a new tool or practice in industry it is important to not only report the evidence of its effectiveness but also the cost of adoption. Given that simulation is perceived as an expensive and non-critical project management activity (Raffo, 2012) makes reporting the cost of conducting an SPSM study even more important.

However except for two studies none of the primary studies reported the effort spent on simulation. Pfahl and Lebsanft (2000a) report 18 months of calendar time for the simulation study and an effort of one person year in consulting and 0.25 person years for the development part. In another study, Pfahl (Pfahl et al., 2004) only reports the calendar time of three months for knowledge elicitation and modelling and four meetings that were held in this period between Fraunhofer IESE and Daimler Chrysler. Shannon (1986) predicted a high cost for simulation as well, *"a practitioner is required to have about 720 hours of formal classroom instruction plus 1440 hours of outside study (more than one man-year of effort)"*.

Given the nature of software development where change is so frequent (technology changes, software development process, environment, and customer requirements, etc.), it is very likely that for simulation as a decision support tool will require adaptation. Expecting practitioners to put in one man-year of effort upfront is very unrealistic (especially under the circumstance that we do not have strong evidence for its usefulness). Furthermore apart from the cost in terms of required tool support, training and effort for development, use and maintenance of simulation models there is the cost of the necessary measurement program that can feed the model with accurate data that should also be acknowledged for an effective process simulation.

6.8. Accessibility of SPSM studies

In conducting this systematic review, it was pivotal to have access to the Proceedings of the "International Software Process Simulation Modeling Workshop" ProSim. However, these proceedings were not available online e.g. the links (at http://www.icsp-conferences.org/icssp2011/previous.html) and the link to the predecessor Prosim (at http://www.prosim.pdx.edu/) were broken. We obtained the proceedings for ProSim 2003–2005 from a personal contact.

6.9. Assessing the evolution of the field

From the point of view of a secondary study that aims to assess and aggregate evidence reported in primary studies, it is very important to understand the contributions of each study. This was particularly difficult in cases of journal articles that were extended from conference articles but did not explicitly acknowledge it, e.g. see the pairs Kusumoto et al. (1997, 2001), Ferreira et al. (2003,

2009), Deissenboeck and Pizka (2007, 2008), Turnu et al. (2004, 2006), Madachy and Khoshnevis (1997) and Madachy (1996), as well as Raffo et al. (2007, 2008).

Furthermore, some studies are remarkably similar, e.g. see the pairs Zhang et al. (2008c,e), Car and Mikac (2002) and Car et al. (2002), as well as Zhang (2009) and Zhang et al. (2009). A methodology to develop a simulation model proposed by Park et al. is reported in three articles which are very similar as well (Park et al., 2007, 2008a,b).

Similarly, Zhang et al. have reported a two phase systematic review in five articles (Zhang et al., 2008a,b,d, 2010, 2011). There is an overlap in the contribution of at least three of these articles (Zhang et al., 2008a,b,d).

From a reporting point of view we therefore recommend to report on reuse of previous models, and extensions made to them to aid other researchers in assessing the contributions made to the field.

In the proceedings of the *"2012 International conference on Software and System Process"* there is an indication that researchers in the SPSM community are trying to ponder on reasons why software process simulation has not had the impact it ought to Münch (2012), Raffo (2012) and Houston (2012). Many of the challenges reported by them are also true in general for all empirical research in SE, e.g. access to practitioners, acquiring credible data, and getting the solutions adopted in industry (Sutton, 2012). Other challenges are perhaps more unique for our research community e.g. SPSM is not sufficiently communicated and is not understood by practitioners (Münch, 2012; Raffo, 2012), and the perceived high cost of SPSM (Raffo, 2012). A more apt reflection on the state of SPSM research is that, *"A retrospective or post-mortem is seldom conducted with SPS so as explain and deepen understanding of a completed project"* (Houston, 2012).

Münch (2012) argues that a major challenge in wide spread adoption of SPSM is that the software industry *"is often not ready for applying process simulation in a beneficial way"*. However, based on the results of this systematic literature review, we can see that it will be difficult to build a case for the use of SPSM in practice without reporting the evidence for its usefulness and the cost of adoption.

The results of this systematic literature review corroborate the claim by Pfahl (2014) that there is no evidence of wide spread adoption and impact of SPSM research on industry. He also challenges if SPSM has a realistic potential to have an impact on industrial practice (Pfahl, 2014). He is not very optimistic about the likelihood of adoption we believe that although he has strong arguments more research is required to make any conclusive statements in this regard. We think that there is a need to identify the problems where the use of simulation can be justified given the high cost of undertaking it and show its utility.

There is a lack of studies evaluating the usefulness of SPSM for training and learning compared to other instructional methods in industrial settings. Pfahl (2014) is however more optimistic about the future of SPSM as a means for learning and training in industry. Based on the results of a systematic literature review (von Wangenheim and Shull, 2009) we do not share the same enthusiasm. von Wangenheim and Shull (2009) considered any study using a game-based (game or simulation) approach for an educational purpose in their review. The studies included in their review were predominantly based on simulation. They found that games are effective to reinforce already acquired knowledge. They also concluded that games should only be used in combination with other methods like lectures with an intended learning outcome. Given that they found that games (including simulation based) only have more *"impact on lower cognitive levels, reinforcing knowledge learned earlier"* it is less likely that we will have learning objectives for practitioners that will justify the use of simulation. For example, we reported a study where a simulation model was used to illustrate the importance of early integration to the practitioners in a large company that develops software intensive products (Ali and Petersen, 2012). However, it is still an open question whether simulation is more effective in getting the message across instead of only static process diagrams, presentations and spreadsheets with graphs showing the potential benefits of early integration and the implications of the current practice.

6.10. Recommendations

For practice and research we provide the following recommendations:

1. When using simulation models for scientific purposes (e.g. assessing the usefulness of test driven development), one has to be sure that the appropriate steps with respect to model validity checking have been conducted. Furthermore, given the limitations in demonstrating the usefulness of simulation beyond replication of reference behaviour, we recommend to not rely on single simulation studies until further evidence for their reliability has been provided. That is, the current state of evidence does not support the common claim that they can replace (controlled) experiments and case studies.
2. From a research perspective, future studies should not just focus on replicating reference behaviour, as many studies have shown that this was successfully achieved. In particular, future studies should go through all necessary steps (building and calibrating models based on a real-world context, establish structural and behavioural model validity, and conduct a series of evaluations of the simulation with respect to its purpose) to significantly drive forward the field of software process simulation. Here, several studies are needed to cover the different purposes more extensively.
3. The success of Step 2 depends very much on complete reporting of how the research study is conducted, in particular with respect to reporting context, data collection, and validity. Important references guiding simulation researchers are Petersen and Wohlin (2009) for describing the context in industrial studies, Runeson and Höst (2009) for case study research, Wohlin et al. (2012) on conducting and reporting experiments. For reporting SPSM studies using continuous simulation a good template is available in Madachy's book (Madachy, 2008).

7. Conclusions

In this paper, we identified literature reporting the application of software process simulation in real-world settings. We identified a total of 87 primary studies. To increase the validity of the results we undertook a number of measures including individually selecting and assessing the articles, conducting pilots in each phase, calculating inter-rater agreement and discussing any case of disagreement.

Articles were assessed based on their reporting for scientific rigor and industrial relevance. Furthermore, we evaluated whether the simulation model's validity was assured. We also determined how studies evaluated simulation against the purpose for which it was used in the real-world. A large majority of the primary studies scored poorly with respect to the rigor and relevance criteria.

With regard to the scoping of real-world simulation studies we conclude that a research gap was identified in relation to simulation of concurrent projects, and the study of long term evolution from a product and organizational perspective. Such projects are of particular interest in the case of large scale software development where several teams work concurrently in software projects for development of an overall system.

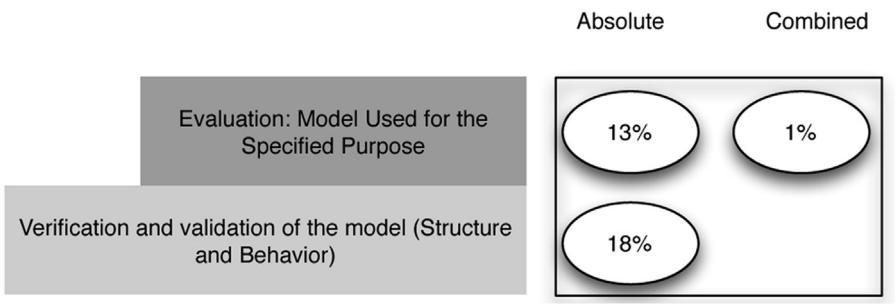

Fig. 3. State of model validation and evaluation.

Fig. 3 provides an overview of achieved validation and evaluation of usefulness. The figure shows that 18% of the primary studies verified the model structure and behaviour. Overall, 13% provided an evaluation and only four of these studies scored high on either rigor or relevance criteria. Furthermore, the evaluation done was at best only static according to the definition from Gorschek et al. (2006), i.e. feedback from practitioners was collected whether the model has the potential to fulfil the intended purpose. Of the overall set, only one article reports having verified structure and behaviour of the model, and evaluated it against the specified purpose.

Based on the results of this systematic literature review that was more extensive in coverage than any of the previously published secondary studies we can draw the following conclusions:

- Despite the large number of industrial applications found in this review, there are no reported cases of the transfer of technology where SPSM was successfully transferred to practitioners in the software industry. Furthermore, there are no studies reporting a long-term use of SPSM in practice. There is no evidence to back the claims of practical adoption and impact on industrial practice (Zhang, 2012; Zhang et al., 2011). This finding supports the position taken by Pfahl (2014) that there is little evidence that process simulation has become an accepted and regularly used tool in industry.
- Based on the reported cost of conducting an SPSM based study (in a few studies), SPSM is not an inexpensive undertaking. Furthermore, without first reporting the cost of adopting SPSM we cannot have any discussion about the cost of *"not simulating"* (Birkhölzer, 2012).
- There is no conclusive evidence to substantiate the claimed benefits of SPSM for any of the proposed purposes. It has been sufficiently argued and claimed in proof-of-concept studies that different simulation approaches *"can be"* used to simulate the software process in varying scopes and that these models *"can be"* used for various purposes. However, the need now is to take the field one step further and provide evidence of these claims by evaluating these research proposals in the real-world. Therefore, while more industry relevant empirical research in SPSM (Sutton, 2012) should be the direction, the goal should be to evaluate the usefulness of SPSM in practice.

In future work, based on our findings, it is important to evaluate simulation against the purposes (e.g. education and training, prediction), which has not been done so far. Future studies should not focus on evaluating the ability of simulation to reproduce reference behaviour, the fact that it is capable to do this is well established in the literature.

Acknowledgments

The authors would like to thank Dietmar Pfahl who provided us with the Proceedings of ProSim Conference from 2003 to 2005. This work has been supported by ELLIIT, a Strategic Area within IT and Mobile Communications, funded by the Swedish Government.

Appendix A. Quality assessment and model validity scores for primary studies

Table A.1
Rigor, relevance and model validity scores for the primary studies.

Reference	Context description	Study design description	Validity discussion	Subjects	Scale	Model validity
Antoniades et al. (2002)	0	0	0	0	0	0.5
Birkhölzer et al. (2005)	0	0	0	0	0	0
Rus et al. (2003)	0	0	0	0	0	0.5
Ruiz et al. (2004)	0	0	0	0	0	0
Raffo et al. (2004)	0	0	0	0	0	0
Katsamakas and Georgantzas (2007)	0	0	0	0	0	1
Wiegner and Nof (1993)	0	0	0	0	0	0.5
Chiang and Menzies (2002)	0	0	0	0	0	0
Bai et al. (2009)	0	0	0	0	0	0
Zhang (2009)	0	0	0	0	0	0.5
Ba and Wu (2012)	0	0	0	0	0	0.5
Dasgupta et al. (2011)	0	0	0	0	0	0
Ferreira et al. (2009)	0	0.5	0	0	0	1
Regnell et al. (2004)	0	0	0.5	0	0	0.5
Gillenson et al. (2011)	0	0.5	0	0	0	0.5
Jian-Hong et al. (2011)	0	0.5	0	0	0	0
Di Penta et al. (2011)	0	0	0.5	0	0	0.5

Table A.1 (*Continued*)

Reference	Context description	Study design description	Validity discussion	Subjects	Scale	Model validity
Crespo and Ruiz (2012)	0	0.5	0	0	0	0.5
Madachy and Khoshnevis (1997)	0	0	0	0	1	0.5
Pfahl et al. (2004)	0	0	0	0	1	0.5
Mizell and Malone (2007b)	0	0	0	0	1	0
Raffo et al. (1999)	0	0	0	0	1	0
Dickmann et al. (2007)	0	0	0	0	1	0
Ruiz et al. (2001)	0	0	0	0	1	1
Wang and Chen (2003)	0	0	0	0	1	0
Lin and Levary (1989)	0	0	0	0	1	0.5
Park et al. (2008b)	0	0	0	0	1	0
Melis et al. (2006)	0	0	0	0	1	1
Raffo et al. (2008)	0	0	0	0	1	0.5
Andersson et al. (2002)	0	0	0	0	1	1
Aranha and Borba (2008)	0	0	0	0	1	0
Meilong et al. (2008)	0	0	0	0	1	0
Wernick and Lehman (1999)	0	0	0	0	1	1
Stallinger (2000)	0	0	0	0	1	0.5
Podnar and Mikac (2001)	0	0	0	0	1	0
Yong and Zhou (2010)	0	0	0	0	1	0.5
Wu et al. (2010)	0	0	0	0	1	0.5
Turnu et al. (2006)	0	0	0	0	1	1
Christie and Staley (2000)	0	0	0	0	1	0
Pfahl and Lebsanft (2000a)	0	0.5	0	0	1	1
Madachy (1995)	0	0	0	0	1	0.5
Mizell and Malone (2007a)	0	0.5	0	0	1	0.5
Pfahl and Lebsanft (2000b)	0	0.5	0	0	1	1
Houston et al. (2001)	0	0.5	0	0	1	1
Al-Emran et al. (2008)	0	0.5	0.5	0	1	0
Abdel-Hamid (1989b)	0	0.5	0	0	1	0.5
Höst et al. (2001)	0	0.5	0.5	0	1	1
Huang et al. (2006)	0	0.5	0	0.5	1	0
Hsueh et al. (2008)	0	0.5	0	0.5	1	0
Kusumoto et al. (2001)	0.5	0	0	0	0	0.5
Tausworthe and Lyu (1994)	0.5	0	0	0	0	0.5
Chen and Wei (2009)	0.5	0	0	0	0	0
Lin (2011)	0.5	0	0	0	0	0
Raffo (2005)	0.5	0	0	0	0	1
Seunghun and Doo-Hwan (2011)	0.5	0	0	0	0	0
Martin and Raffo (2001)	0.5	0.5	0	0	0	0.5
Car et al. (2002)	0.5	0	0	0	1	0.5
Abdel-Hamid (1990)	0.5	0	0	0	1	0.5
Abdel-Hamid (1993)	0.5	0	0	0	1	0.5
Abdel-Hamid (1988b)	0.5	0	0	0	1	0.5
Abdel-Hamid and Leidy (1991)	0.5	0	0	0	1	0.5
McCall et al. (1979)	0.5	0	0	0	1	0.5
Abdel-Hamid (1989a)	0.5	0	0	0	1	0.5
Mizuno et al. (2001)	0.5	0	0	0	1	0
Mizuno et al. (1997)	0.5	0	0	0	1	0.5
Nakatani and Nishida (1992)	0.5	0	0	0	1	0.5
Ramil and Smith (2002)	0.5	0	0	0	1	0.5
Anderson et al. (2012)	0.5	0	0	0	1	0.5
Psaroudakis and Eberhardt (2011)	0.5	0	0	0	1	0.5
Chatters et al. (2000)	0.5	0.5	0	0	1	1
Setamanit and Raffo (2008)	0.5	0.5	0	0	1	0.5
Paikari et al. (2012)	0.5	0.5	0	0	1	0
Farshchi et al. (2012)	0.5	0.5	0	0	1	0.5
Deissenboeck and Pizka (2008)	0.5	0.5	0.5	0	1	0.5
Spasic and Onggo (2012)	0.5	1	0	0	1	0.5
Houston (2006)	0.5	0	0	1	1	0.5
Al-Emran et al. (2010)	0.5	0.5	0.5	1	1	0
Zhang et al. (2008e)	0.5	0	0	0	0	0
Li et al. (2006)	0.5	0	0	0	1	0.5
Junjie et al. (2012)	1	0.5	0.5	0	0	0.5
Shen et al. (2005)	1	0	0	0	1	0.5
Abdel-Hamid (1988a)	1	0	0	0	1	0.5
Antoniol et al. (2004)	1	0.5	0	0	1	0.5
Lin et al. (1997)	1	0.5	0	0	1	1
Rahmandad and Weiss (2009)	1	0.5	0	0	1	1
Car et al. (2010)	1	0.5	0	0	1	1
Ghosh and Wei (2010)	1	1	0.5	0	1	0.5

References

Ahmed, R., Hall, T., Wernick, P., Robinson, S., 2005. Evaluating a rapid simulation modelling process (RSMP) through controlled experiments. In: Proceedings of the International Symposium on Empirical Software Engineering, Piscataway, NJ, USA.

Ahmed, R., Hall, T., Wernick, P., 2003. A proposed framework for evaluating software process simulation models. In: Proceedings of the International Workshop on Software Process Simulation and Modeling (ProSim).

Ahmed, R., Hall, T., Wernick, P., Robinson, S., Shah, M., 2008. Software process simulation modelling: a survey of practice. J. Simulat. 2 (2), 91–102.

Ali, N.b., Petersen, K., 2012. A consolidated process for software process simulation: state of the art and industry experience. In: Proceedings of the 38th IEEE EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), pp. 327–336.

Ali, N.b., Petersen, K., Wohlin, C., 2014. Supporting Information for the Systematic Literature Review. http://www.bth.se/com/nal.nsf/pages/spsm

Bai, X., Zhang, H., Huang, L., 2011. Empirical research in software process modeling: a systematic literature review. In: International Symposium on Empirical Software Engineering and Measurement, ESEM, pp. 339–342.

Balci, O.,1990. Guidelines for successful simluation studies. In: Proceedings of the Winter Simulation Conference. IEEE Press, pp. 25–32.

Basili, V., 1985. Quantitative evaluation of software methodology. In: Proceedings of the First Pan Pacific Computer Conference, Melbourne, Australia.

Birkhölzer, T., 2012. Software process simulation is simulation too – what can be learned from other domains of simulation? In: 2012 International Conference on Software and System Process (ICSSP), pp. 223–225, http://dx.doi.org/10.1109/ICSSP.2012.6225972.

Car, Z., Mikac, B., 2002. A method for modeling and evaluating software maintenance process performances. In: Proceedings of the Sixth European Conference on Software Maintenance and Reengineering, pp. 15–23.

de Almeida Biolchini, J., Mian, P., Natali, A., Conte, T., Travassos, G., 2007. Scientific research ontology to support systematic review in software engineering. Adv. Eng. Informat. 21 (2), 133–151.

Deissenboeck, F., Pizka, M.,2007. The economic impact of software process variations. In: Proceedings of the 2007 International Conference on Software Process. Springer-Verlag, pp. 259–271.

Dybå, T., 2013. Contextualizing empirical evidence. IEEE Softw. 30 (1), 81–83.

Ferreira, S., Collofello, J., Shunk, D., Mackulak, G., Wolfe, P., 2003. Utilization of process modeling and simulation in understanding the effects of requirements volatility in software development. In: Proceedings of the International Workshop on Software Process Simulation and Modeling (ProSim).

França, B.B.N.d., Travassos, G.H., 2013. Are we prepared for simulation based studies in software engineering yet? CLEI Electron. J. 16 (1), 9.

Gorschek, T., Wohlin, C., Garre, P., Larsson, S., 2006. A model for technology transfer in practice. IEEE Softw. 23 (6), 88–95.

Houston, D.,2012. Research and practice reciprocity in software process simulation. In: International Conference on Software and System Process (ICSSP). IEEE, pp. 219–220.

ISO/IEC, 2003–2006. ISO/IEC 15504 information technology – process assessment (Parts 1–5).

Ivarsson, M., Gorschek, T., 2010. A method for evaluating rigor and industrial relevance of technology evaluations. Empir. Softw. Eng. 16 (3), 365–395.

Kellner, M.I., Madachy, R.J., Raffo, D.M., 1999. Software process simulation modeling: why? what? how? J. Syst. Softw. 46, 91–105.

Khan, K.S., ter Riet, G., Glanville, J., Sowden, A.J., Kleijnen, J., 2001. Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for Carrying Out or Commissioning Reviews, vol. 4., 2nd ed. NHS Centre for Reviews and Dissemination, University of York, York, United Kingdom.

Kitchenham, B., Brereton, P., Budgen, D., 2012. Mapping study completeness and reliability – a case study. In: Proceedings of the 16th International Conference on Evaluation Assessment in Software Engineering (EASE 2012), pp. 126–135.

Kitchenham, B., Brereton, P., Li, Z., Budgen, D., Burn, A., 2011. Repeatability of systematic literature reviews. In: Proceedings of the 15th Annual Conference on Evaluation Assessment in Software Engineering (EASE), pp. 46–55.

Kitchenham, B., Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering. Tech. Rep. EBSE 2007-001, Keele University and Durham University Joint Report.

Kitchenham, B., 2010. What's up with software metrics? – A preliminary mapping study. J. Syst. Softw. 83 (1), 37–51.

Kusumoto, S., Mizuno, O., Kikuno, T., Hirayama, Y., Takagi, Y., Sakamoto, K.,1997. A new software project simulator based on generalized stochastic Petri-net. In: Proceedings of the 19th International Conference on Software Engineering. ACM, pp. 293–302.

Law, A.M., 2001. How to build valid and credible simulation models. In: Proceeding of the 2001 Winter Simulation Conference, pp. 22–29.

Liu, D., Wang, Q., Xiao, J.,2009. The role of software process simulation modeling in software risk management: a systematic review. In: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement. IEEE Computer Society, pp. 302–311.

Münch, J., 2012. Evolving process simulators by using validated learning. In: 2012 International Conference on Software and System Process (ICSSP), pp. 226–227.

Madachy, R., 2002. Simulation. In: Marciniak, J.J. (Ed.), Encyclopedia of Software Engineering. John Wiley & Sons, Inc., Hoboken, NJ, USA, http://dx.doi.org/10.1002/0471028959.

Madachy, R.J.,1996. System dynamics modeling of an inspection-based process. In: Proceedings of the 18th International Conference on Software Engineering. IEEE, pp. 376–386.

Madachy, R.J., 2008. Software Process Dynamics. Wiley-IEEE Press, Hoboken, New Jersey.

Mujtaba, S., Feldt, R., Petersen, K., 2010. Waste and lead time reduction in a software product customization process with value stream maps. In: Proceedings of the 21st Australian Software Engineering Conference (ASWEC), pp. 139–148.

Park, S.H., Choi, K.S., Yoon, K., Bae, D.-H.,2007. Deriving software process simulation model from SPEM-based software process model. In: Proceedings of the 14th Asia-Pacific Software Engineering Conference. IEEE, pp. 382–389.

Park, S., Kim, H., Kang, D., Bae, D.-H.,2008a. Developing a simulation model using a SPEM-based process model and analytical models. In: Proceedings of the 4th International Workshop CIAO! and 4th International Workshop EOMAS. Springer, pp. 164–178.

Petersen, K., Ali, N.,2011. Identifying strategies for study selection in systematic reviews and maps. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement (ESEM). IEEE, pp. 351–354.

Petersen, K., Wohlin, C., 2009. Context in industrial software engineering research. In: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement, pp. 401–404.

Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M., 2008. Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, pp. 71–80.

Pfahl, D., 2014. Process simulation – a tool for software project managers? In: Ruhe, G., Wohlin, C. (Eds.), Software Project Management in a Changing World. Springer-Verlag Berlin Heidelberg, New York, USA.

Raffo, D., 2012. Process simulation will soon come of age: where's the party? In: 2012 International Conference on Software and System Process (ICSSP), p. 231, http://dx.doi.org/10.1109/ICSSP.2012.6225975.

Raffo, D.M., Ferguson, R., Setamanit, S.-O., Sethanandha, B.D.,2007. Evaluating the impact of the QuARS requirements analysis tool using simulation. In: Proceedings of the International Conference on Software Process, ICSP'07. Springer-Verlag, Berlin/Heidelberg, pp. 307–319.

Ruiz, M., Ramos, I., Toro, M., 2002. A dynamic integrated framework for software process improvement. Softw. Qual. J. 10 (2), 181–194.

Runeson, P., Höst, M., 2009. Guidelines for conducting and reporting case study research in software engineering. Empir. Softw. Eng. 14 (2), 131–164.

Rus, I., Neu, H., Münch, J., 2003. A systematic methodology for developing discrete event simulation models of software development processes. In: Proceedings of the International Workshop on Software Process Simulation Modeling (ProSim), Citeseer.

Shannon, R.E., 1986. Intelligent simulation environments. In: Proceedings of the Conference on Intelligent Simulation Environments, pp. 150–156.

Shull, F., Singer, J., Sjøberg, D., 2007. Guide to Advanced Empirical Software Engineering. Springer.

Slawson, D.C., 1997. How to read a paper: the basics of evidence based medicine. BMJ 315 (7112), 891.

Sutton, S., 2012. Advancing process modeling, simulation, and analytics in practice. In: International Conference on Software and System Process (ICSSP), pp. 221–222.

Team, C.P., 2006. CMMI for Development, Version 1.2. Carnegie Mellon, Software Engineering Institute, Pittsburgh, PA.

Turnu, I., Melis, M., Cau, A., Marchesi, M., Setzu, A.,2004. Introducing TDD on a free libre open source software project: a simulation experiment. In: Proceedings of the 2004 Workshop on Quantitative Techniques for Software Agile Process. ACM, pp. 59–65.

von Wangenheim, C., Shull, F., 2009. To game or not to game? IEEE Softw., 92–94.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2012. Experimentation in Software Engineering. Springer-Verlag Berlin Heidelberg, New York, USA.

Wohlin, C., Runeson, P., da Mota Silveira Neto, P.A., Engström, E., do Carmo Machado, I., de Almeida, E.S., 2013. On the reliability of mapping studies in software engineering. J. Syst. Softw. 86 (10), 2594–2610.

Wohlin, C., 1991. Software reliability and performance modelling for telecommunication systems. Ph.D. thesis.

Zhang, H., Kitchenham, B., Pfahl, D., 2008a. Reflections on 10 Years of Software Process Simulation Modeling: A Systematic Review. Springer, Berlin, Heidelberg, Germany.

Zhang, H., Kitchenham, B., Pfahl, D.,2008b. Software process simulation modeling: facts, trends and directions. In: Proceedings of the 15th Asia-Pacific Software Engineering Conference (APSEC). IEEE, pp. 59–66.

Zhang, H., Jeffery, R., Zhu, L.,2008c. Hybrid modeling of test-and-fix processes in incremental development. In: Proceedings of the International Conference on Software Process (ICSP). Springer, pp. 333–344.

Zhang, H., Jeffery, R., Houston, D., Huang, L., Zhu, L., 2011. Impact of process simulation on software practice: an initial report. In: Proceedings of the 33rd International Conference on Software Engineering (ICSE), pp. 1046–1056.

Zhang, H., Kitchenham, B., Jeffery, R.,2009. Qualitative vs. quantitative software process simulation modeling: conversion and comparison. In: Proceedings of the Australian Software Engineering Conference. IEEE, pp. 345–354.

Zhang, H., Kitchenham, B., Pfahl, D.,2010. Software process simulation modeling: an extended systematic review. In: Proceedings of the International Conference on Software Process (ICSP). Springer, pp. 309–320.

Zhang, H., Kitchenham, B., Pfahl, D.,2008d. Software process simulation over the past decade: trends discovery from a systematic review. In: Proceedings of the

Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement. ACM, Kaiserslautern, Germany, pp. 345–347.
Zhang, H.,2012. Special panel: software process simulation – at a crossroads? In: Proceedings of the International Conference on Software and System Process (ICSSP). IEEE, pp. 215–216.

Systematic literature review

Abdel-Hamid, T., 1988a. The economics of software quality assurance: a simulation-based case study. MIS Quart. 12 (3), 395–411.
Abdel-Hamid, T., 1988b. Understanding the "90% syndrome" in software project management: a simulation-based case study. J. Syst. Softw. 8 (4), 319–330.
Abdel-Hamid, T., 1989a. The dynamics of software project staffing: a system dynamics based simulation approach. IEEE Trans. Softw. Eng. 15 (2), 109–119.
Abdel-Hamid, T., 1990. On the utility of historical project statistics for cost and schedule estimation: results from a simulation-based case study. J. Syst. Softw. 13 (1), 71–82.
Abdel-Hamid, T., Leidy, F., 1991. An expert simulator for allocating the quality assurance effort in software development. Simulation 56 (4), 233–240.
Abdel-Hamid, T., 1993. A multiproject perspective of single-project dynamics. J. Syst. Softw. 22 (3), 151–165.
Abdel-Hamid, T.K., 1989b. A study of staff turnover, acquisition, and assimilation and their impact on software development cost and schedule. J. Manage. Inform. Syst. 6 (1), 21–40.
Al-Emran, A., Kapur, P., Pfahl, D., Ruhe, G.,2008. Simulating worst case scenarios and analyzing their combined effect in operational release planning. In: Proceedings of the International Conference on Software Process, ICSP, vol. 5007. Springer, p. 269.
Al-Emran, A., Kapur, P., Pfahl, D., Ruhe, G., 2010. Studying the impact of uncertainty in operational release planning – an integrated method and its initial evaluation. Inform. Softw. Technol. 52 (4), 446–461.
Anderson, D.J., Concas, G., Lunesu, M.I., Marchesi, M., Zhang, H., 2012. A comparative study of Scrum and Kanban approaches on a real case study using simulation. In: 13th International Conference on Agile Software Development, XP 2012, vol. 111. LNBIP, May 21, 2012–May 25, 2012. Springer Verlag, pp. 123–137.
Andersson, C., Karlsson, L., Nedstam, J., Höst, M., Nilsson, B.,2002. Understanding software processes through system dynamics simulation: a case study. In: Proceedings of the Ninth Annual IEEE International Conference and Workshop on the Engineering of Computer-Based Systems. IEEE, pp. 41–48.
Antoniades, I., Stamelos, I., Angelis, L., Bleris, G., 2002. A novel simulation model for the development process of open source software projects. Softw. Process: Improv. Pract. 7 (3/4), 173–188.
Antoniol, G., Cimitile, A., Di Lucca, G., Di Penta, M., 2004. Assessing staffing needs for a software maintenance project through queuing simulation. IEEE Trans. Softw. Eng. 30 (1), 43–58.
Aranha, E., Borba, P.,2008. Using process simulation to assess the test design effort reduction of a model-based testing approach. In: Proceedings of the International Conference on Software process (ICSP). Springer-Verlag, pp. 282–293.
Ba, J., Wu, S.,2012. A dynamic defect prediction model. In: Proceedings of 2012 8th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications, MESA. IEEE Computer Society, pp. 252–256.
Bai, X., Huang, L., Koolmanojwong, S.,2009. Incremental process modeling through stakeholder based hybrid process simulation. In: Proceedings of the International Conference on Software Process (ICSP). Springer-Verlag, Berlin, Germany, pp. 280–292.
Birkhölzer, T., Dickmann, C., Vaupel, J., Dantas, L., 2005. An interactive software management simulator based on the CMMI framework. Softw. Process: Improv. Pract. 10 (3), 327–340.
Cao, L., Ramesh, B., Abdel-Hamid, T., 2010. Modeling dynamics in agile software development. ACM Trans. Manage. Inform. Syst. 1 (1), http://dx.doi.org/10.1145/1877725.1877730.
Car, Z., Mikac, B., Sinkovic, V., 2002. A simulation method for telecommunication software maintenance process. In: Proceedings of the 20th IASTED International Conference Applied Informatics, AI 2002, pp. 69–74.
Chatters, B., Lehman, M., Ramil, J., Wernick, P., 2000. Modelling a software evolution process: a long-term case study. Softw. Process: Improv. Pract. 5 (2/3), 91–102.
Chen, Y.M., Wei, C.-W., 2009. Multiagent approach to solve project team work allocation problems. Int. J. Prod. Res. 47 (13), 3453–3470.
Chiang, E., Menzies, T., 2002. Simulations for very early lifecycle quality evaluations. Softw. Process: Improv. Pract. 7 (3/4), 141–159.
Christie, A.M., Staley, M.J., 2000. Organizational and social simulation of a software requirements development process. Softw. Process: Improv. Pract. 5 (2/3), 103–110.
Crespo, D., Ruiz, M., 2012. Decision making support in CMMI process areas using multiparadigm simulation modeling. In: Proceedings of the 2012 Winter Simulation Conference (WSC 2012), December 9–12, 2012. IEEE, p. 12.
Dasgupta, J., Sahoo, G., Mohanty, R.P., 2011. Monte Carlo simulation based estimations: case from a global outsourcing company. In: Proceedings of the 1st International Technology Management Conference, ITMC 2011, June 27, 2011–June 30, 2011. IEEE Computer Society, pp. 619–624.
Deissenboeck, F., Pizka, M., 2008. Probabilistic analysis of process economics. Softw. Process: Improv. Pract. 13 (1), 5–17.
Di Penta, M., Harman, M., Antoniol, G., 2011. The use of search-based optimization techniques to schedule and staff software projects: an approach and an empirical study. Softw. Pract. Exp. 41 (5), 495–519.
Dickmann, C., Klein, H., Birkhölzer, T., Fietz, W., Vaupel, J., Meyer, L.,2007. Deriving a valid process simulation from real world experiences. In: Proceedings of the International Conference on Software Process. Springer, pp. 272–282.
Farshchi, M., Jusoh, Y.Y., Murad, M.A.A., 2012. Impact of personnel factors on the recovery of delayed software projects: a system dynamics approach. Comput. Sci. Inform. Syst. 9 (2), 627–651.
Ferreira, S., Collofello, J., Shunk, D., Mackulak, G., 2009. Understanding the effects of requirements volatility in software engineering by using analytical modeling and software process simulation. J. Syst. Softw. 82 (10), 1568–1577.
Ghosh, J., Concurrent, 2010. overlapping development and the dynamic system analysis of a software project. IEEE Trans. Eng. Manage. 57 (2), 270–287.
Gillenson, M.L., Racer, M.J., Richardson, S.M., Zhang, X., 2011. Engaging testers early and throughout the software development process: six models and a simulation study. J. Inform. Technol. Manage. 22 (1), 8–27.
Höst, M., Regnell, B., Natt och Dag, J., Nedstam, J., Nyberg, C., 2001. Exploring bottlenecks in market-driven requirements management processes with discrete event simulation. J. Syst. Softw. 59 (3), 323–332.
Houston, D., 2006. An experience in facilitating process improvement with an integration problem reporting process simulation. Softw. Process: Improv. Pract. 11 (4), 361–371.
Houston, D.X., Mackulak, G.T., Collofello, J.S., 2001. Stochastic simulation of risk factor potential effects for software development risk management. J. Syst. Softw. 59 (3), 247–257.
Hsueh, N., Shen, W., Yang, Z., Yang, D., 2008. Applying UML and software simulation for process definition, verification, and validation. Inform. Softw. Technol. 50 (9), 897–911.
Huang, L., Boehm, B., Hu, H., Ge, J., Lü, J., Qian, C.,2006. Applying the value/Petri process to ERP software development in China. In: Proceedings of the 28th International Conference on Software Engineering. ACM, pp. 502–511.
Jian-Hong, H., Xiaoing, B., Qi, L., Daode, X., 2011. A fuzzy-ECM approach to estimate software project schedule under uncertainties. In: 2011 Ninth IEEE International Symposium on Parallel and Distributed Processing with Applications Workshops (ISPAW), pp. 316–321.
Junjie, W., Juan, L., Qing, W., He, Z., Haitao, W., 2012. A simulation approach for impact analysis of requirement volatility considering dependency change. In: Requirements Engineering: Foundation for Software Quality, Proceedings 18th International Working Conference, REFSQ 2012, March 19–22, 2012. Springer-Verlag, pp. 59–76.
Katsamakas, E., Georgantzas, N.,2007. Why most open source development projects do not succeed? In: Proceedings of the First International Workshop on Emerging Trends in FLOSS Research and Development. IEEE, p. 3.
Kusumoto, S., Mizuno, O., Kikuno, T., Hirayama, Y., Takagi, Y., Sakamoto, K., 2001. Software project simulator for effective process improvement. J. Inform. Process. Soc. Jpn. 42 (3), 396–408.
Li, X.M., Wang, Y.L., Sun, L.Y., Li, L., 2006. Modeling uncertainties involved with software development with a stochastic Petri net. Expert Syst. 23 (5), 302–312.
Lin, C.T., 2011. Analyzing the effect of imperfect debugging on software fault detection and correction processes via a simulation framework. Math. Comput. Model. 54 (11/12), 3046–3064.
Lin, C., Levary, R., 1989. Computer-aided software development process design. IEEE Trans. Softw. Eng. 15 (9), 1025–1037.
Lin, C.Y., Abdel-Hamid, T., Sherif, J.S., 1997. Software-engineering process simulation model (SEPS). J. Syst. Softw. 38 (3), 263–277.
Madachy, R.J., 1995. Knowledge-based risk assessment and cost estimation. Automat. Softw. Eng. 2 (3), 219–230.
Madachy, R.J., Khoshnevis, B., 1997. Dynamic simulation modeling of an inspection-based software lifecycle process. Simulation 69 (1), 35–47.
Martin, R., Raffo, D., 2001. Application of a hybrid process simulation model to a software development project. J. Syst. Softw. 59 (3), 237–246.
McCall, J.A., Wong, G., Stone, A.,1979. A simulation modeling approach to understanding the software development process. In: Fourth Annual Software Engineering Workshop. Goddard Space Flight Center Greenbelt, Maryland, pp. 250–273.
Meilong, X., Congdong, L., Jie, C., 2008. System dynamics simulation to support decision making in software development project. In: Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM, pp. 1–4.
Melis, M., Turnu, I., Cau, A., Concas, G., 2006. Evaluating the impact of test-first programming and pair programming through software process simulation. Softw. Process: Improv. Pract. 11 (4), 345–360.
Mizell, C., Malone, L., 2007a. A software development simulation model of a spiral process. Int. J. Softw. Eng. 2 (2).
Mizell, C., Malone, L., 2007b. A project management approach to using simulation for cost estimation on large, complex software development projects. Eng. Manage. J. 19 (4), 28–34.
Mizuno, O., Kusumoto, S., Kikuno, T., Takagi, Y., Sakamoto, K.,1997. Estimating the number of faults using simulator based on generalized stochastic Petri-net model. In: Proceedings of the Sixth Asian Test Symposium (ATS'97). IEEE, pp. 269–274.
Mizuno, O., Shimoda, D., Kikuno, T., Takagi, Y., 2001. Enhancing software project simulator toward risk prediction with cost estimation capability. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 84 (11), 2812–2821.

Nakatani, M., Nishida, S., 1992. Trouble communication model in a software development project. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 75 (2), 196–206.
Paikari, E., Ruhe, G., Southekel, P.H., 2012. Simulation-based decision support for bringing a project back on track: the case of RUP-based software construction. In: 2012 International Conference on Software and System Process, ICSSP 2012 – Proceedings, June 2, 2012–June 3, 2012. Association for Computing Machinery, pp. 13–22.
Park, S., Kim, H., Kang, D., Bae, D., 2008b. Developing a software process simulation model using SPEM and analytical models. Int. J. Simulat. Process Model. 4 (3), 223–236.
Pfahl, D., Lebsanft, K., 2000a. Knowledge acquisition and process guidance for building system dynamics simulation models: an experience report from software industry. Int. J. Softw. Eng. Knowl. Eng. 10 (4), 487–510.
Pfahl, D., Lebsanft, K., 2000b. Using simulation to analyse the impact of software requirement volatility on project performance. Inform. Softw. Technol. 42 (14), 1001–1008.
Pfahl, D., Stupperich, M., Krivobokova, T., 2004. PL-SIM: a generic simulation model for studying strategic SPI in the automotive industry. In: Proceedings of the International Workshop on Software Process Simulation and Modeling (ProSim), pp. 149–158.
Podnar, I., Mikac, B.,2001. Software maintenance process analysis using discrete-event simulation. In: Proceedings of the Fifth European Conference on Software Maintenance and Reengineering. IEEE, pp. 192–195.
Psaroudakis, J.E., Eberhardt, A., 2011. A discrete event simulation model to evaluate changes to a software project delivery process. In: 2011 IEEE 13th Conference on Commerce and Enterprise Computing, September 5–7, 2011. IEEE Computer Society, pp. 113–120, http://dx.doi.org/10.1109/cec.2011.19.
Raffo, D., Ferguson, R., Setamanit, S., Sethanandha, B., 2008. Evaluating the impact of requirements analysis tools using simulation. Softw. Process: Improv. Pract. 13 (1), 63–73.
Raffo, D., Vandeville, J., Martin, R., 1999. Software process simulation to achieve higher CMM levels. J. Syst. Softw. 46 (2), 163–172.
Raffo, D.M., Nayak, U., Setamanit, S.-o., Sullivan, P., Wakeland, W., 2004. Using software process simulation to assess the impact of IV&V activities. In: Proceedings of the 5th International Workshop on Software Process Simulation and Modeling (ProSim), pp. 197–205.
Raffo, D.M., 2005. Software project management using PROMPT: a hybrid metrics, modeling and utility framework. Inform. Softw. Technol. 47 (15), 1009–1017.
Rahmandad, H., Weiss, D.M., 2009. Dynamics of concurrent software development. Syst. Dyn. Rev. 25 (3), 224–249, http://dx.doi.org/10.1002/sdr.425.
Ramil, J.F., Smith, N., 2002. Qualitative simulation of models of software evolution. Softw. Process: Improv. Pract. 7 (3/4), 95–112.
Regnell, B., Ljungquist, B., Thelin, T., Karlsson, L., 2004. Investigation of requirements selection quality in market-driven software processes using an open source discrete event simulation framework. In: Proceedings of the 5th International Workshop on Software Process Simulation and Modeling (ProSim), Stevenage, UK, pp. 84–93.
Ruiz, M., Ramos, I., Toro, M., 2001. A simplified model of software project dynamics. J. Syst. Softw. 59 (3), 299–309.
Ruiz, M., Ramos, I., Toro, M.,2004. Using dynamic modeling and simulation to improve the cots software process. In: Proceedings of the 5th International Conference on Product Focused Software Process Improvement. Springer, pp. 568–581.
Rus, I., Shull, F., Donzelli, P.,2003. Decision support for using software inspections. In: Proceedings of the 28th Annual NASA GoddardSoftware Engineering Workshop. IEEE, pp. 3–11.
Setamanit, S.-O., Raffo, D.,2008. Identifying key success factors for globally distributed software development project using simulation: a case study. In: Proceedings of the International Conference on Software Process. Springer-Verlag, pp. 320–332.
Seunghun, P., Doo-Hwan, B., 2011. An approach to analyzing the software process change impact using process slicing and simulation. J. Syst. Softw. 84 (4), 528–543.
Shen, J., Changchien, S., Lin, T., 2005. A Petri-net based modeling approach to concurrent software engineering tasks. J. Inform. Sci. Eng. 21 (4), 767–795.
Spasic, B., Onggo, B.S.S., 2012. Agent-based simulation of the software development process: a case study at AVL. In: Proceedings of the 2012 Winter Simulation Conference (WSC 2012), December 9–12, 2012. IEEE, p. 11.
Stallinger, F., 2000. Software process simulation to support ISO/IEC 15504 based software process improvement. Softw. Process: Improv. Pract. 5 (2/3), 197–209.
Tausworthe, R., Lyu, M.,1994. A generalized software reliability process simulation technique and tool. In: Proceedings of the 5th International Symposium on Software Reliability Engineering, 1994. IEEE, pp. 264–273.
Turnu, I., Melis, M., Cau, A., Setzu, A., Concas, G., Mannaro, K., 2006. Modeling and simulation of open source development using an agile practice. J. Syst. Architect. 52 (11), 610–618.
Wang, Y., Chen, Y., 2003. An experience of modeling and simulation in support of CMMI process. In: Proceedings of the 15th European Simulation Symposium, pp. 297–302.
Wernick, P., Lehman, M., 1999. Software process white box modelling for FEAST/1. J. Syst. Softw. 46 (2), 193–201.
Wiegner, R.T., Nof, S.Y., 1993. The software product feedback flow model for development planning. Inform. Softw. Technol. 35 (8), 427–438.
Wu, D., Song, H., Li, M., Cai, C., Li, J.,2010. Modeling risk factors dependence using copula method for assessing software schedule risk. In: Proceedings of the 2nd International Conference on Software Engineering and Data Mining (SEDM). IEEE, pp. 571–574.
Yong, Y., Zhou, B.,2010. Software process deviation threshold analysis by system dynamics. In: Proceedings of the 2nd IEEE International Conference on Information Management and Engineering (ICIME). IEEE, pp. 121–125.
Zhang, H., Jeffery, R., Zhu, L.,2008e. Investigating test-and-fix processes of incremental development using hybrid process simulation. In: Proceedings of the 6th International Workshop on Software Quality. ACM, pp. 23–28.
Zhang, H.,2009. Investigating the gap between quantitative and qualitative/semi-quantitative software process simulation models: an explorative study. In: Proceedings of the International Conference on Software Process (ICSP), vol. 5543. Springer, p. 198.

Nauman Bin Ali is a Ph.D. student at Blekinge Institute of Technology. He is involved in empirical research in the field of software engineering. His research interests are software process simulation, software quality, and lean software development. He received his Tek. Lic. (2013) and MSc. (2010) in software engineering from Blekinge Institute of Technology, Sweden.

Kai Petersen is an assistant professor at Blekinge Institute of Technology (BTH), Sweden. He received his Ph.D. from BTH in 2010. His research focuses on software processes, software metrics, lean and agile software development, quality assurance, and software security in close collaboration with industry partners. Kai has authored over 30 articles in international journals and conferences, and has been the industry chair of REFSQ 2013.

Claes Wohlin is a professor of software engineering and dean for the Faculty of Computing at Blekinge Institute of Technology, Sweden. He has previously held professor chairs at the universities in Lund and Linköping. His research interests include empirical methods in software engineering, software metrics, software quality, and requirements engineering. Wohlin received a Ph.D. in communication systems from Lund University. He is Editor-in-Chief of Information and Software Technology journal. He is a member of the Royal Swedish Academy of Engineering Sciences and a senior member of IEEE.