A Status Report on the Evaluation of Variability Management Approaches

Lianping Chen, Muhammad Ali Babar, Ciaran Cawley
Lero, the Irish Software Engineering Research Centre, Limerick
{Lianping.chen, malibaba, ciaran.cawley}@lero.ie

OBJECTIVE – The objective of this study is to review the status of evaluation of Variability Management (VM) approaches in Software Product Line Engineering (SPLE).

METHOD – We have carried out a systematic review of studies of VM in SPLE reported in any publication venue and published before September 2007.

RESULTS – 97 papers were selected according to our inclusion and exclusion criteria. The selected papers appeared in 56 publication venues and the dominance of well-known publication venues of SPLE is not very clear. Only a small portion of the identified approaches were evaluated using rigorous scientific methods. A detailed investigation of the studies employing empirical research methods revealed significant quality deficiencies in various quality assessment aspects. Even more disappointing finding is that the trends of scientific evaluation of VM approaches do not appear to be improving based on the analysis of the data extracted from the reviewed studied.

CONCLUSIONS – The status of evaluation of VM approaches in SPLE is poor. Most of the approaches proposed are amenable to empirical evaluation, but the available evidence is sparse and the quality of the presented evidence is very low. The findings highlight the areas of improvement for rigorous evaluation of VM approaches.

Keywords: Software product line, variability management, systematic literature reviews, empirical studies

1. INTRODUCTION

Software Product Line Engineering (SPLE) intends to develop software-intensive systems using platforms and mass customisation (Bosch 2000; Clements and Northrop 2002). This is achieved through the identification and management of commonalities and variations in a set of systems' artefacts such as requirements, architectures, components, and test cases. A Software Product Line (SPL) is a set of software-intensive systems that share a common and managed set of features that satisfy the specific needs of a particular market segment and that are developed from a common set of core assets in a prescribed way (Bosch 2000). In the product line approach, variability provides the required flexibility for product differentiation and diversification. Variability refers to the ability of an artefact to be configured, customized, extended, or changed for use in a specific context (Bachmann and Clements 2005). Variability Management (VM) encompasses the activities of explicitly representing variability in software artefacts throughout the lifecycle, managing dependencies among different variabilities and supporting the instantiations of those variabilities (Schmid and John 2004). It involves extremely complex and challenging tasks, which needs to be supported by appropriate approaches, techniques, and tools (Bosch, Florijn et al. 2002; Sinnema and Deelstra 2007). Systematically identifying and appropriately managing variabilities among different systems of a family are the key characteristics that distinguish SPLE from other reuse-based software development approaches (Bosch, Florijn et al. 2002).

Given such a vital role of VM in SPLE, there has been a great deal of research in this area of SPLE. Many diverse approaches and tools have been developed with the basic aim of supporting (or automating) various tasks involved in VM at different stages of a product line's life. Like any other software development technology, rigorous evaluation of VM approaches is also very important in order to provide sufficient evidence to support the claimed benefits of the proposed approach. Such evidence can play an important role in transferring the research outcomes into industrial practices (Kitchenham, Dyba et al. 2004). However, there has been no effort to systematically collect and analyse existing evidence on the nature and quality of evaluation of VM approaches reported in the literature. Hence, we decided to conduct a Systematic Literature Review (SLR) or Systematic Review (SR) of the literature on VM in SPLE in order to provide a snapshot of the state-of-the-art with respect to evaluation of proposed approaches. The specific research questions that motivated our study are:

- How have the variability management approaches in SPLE been evaluated?

- What is the quality of the reported evaluations of the variability management approaches?

There have been a few efforts to survey the literature on VM in software product lines (Svahnberg, van Gurp et al. 2005; Sinnema and Deelstra 2007). However, these studies were aimed at very specific elements of VM (i.e., modelling (Sinnema and Deelstra 2007) and realization mechanisms (Svahnberg, van Gurp et al. 2005)). Additionally, our research has completely different goals as stated before and we used a systematic and rigorous approach to identifying and selecting the reviewed primary studies. Our study is based on a systematic search of publications from various data sources and follows a pre-defined protocol during the whole study process. None of the previous reviews describe a systematic selection process of the reviewed studies; nor do they focus on the type and strength of the evidence provided to support the VM approaches.

Section 2 describes the methodology of the systematic review which follows the guidelines as presented in (Kitchenham and Charters 2007). Section 3 presents the results, while Section 4 discusses the results. Finally, Section 5 concludes the paper by discussing our ongoing work on analysing the data gathered from the primary studies reviewed in this study.

2. REVIEW METHOD

As we mentioned, this study has been carried out according to the SR methodology described in (Kitchenham and Charters 2007). Since many recently published studies have described the methodology, logistics, benefits, and limitations of SRs in software engineering, we limit our discussion about the methodological aspects of the reported research. However, we followed the stages and steps recommended in Kitchenham's guidelines updated in 2007. A detailed methodological description of this SR has been provided in a technical report (Chen, Babar et al. 2009).

2.1. Planning the review

In the planning stage, we identified the need for the review, specified research questions, developed a review protocol, and evaluated the review protocol. The protocol was piloted by the researchers. Several issues were identified during the pilot, and we fixed them accordingly. For example, we found that some relevant papers were not present in the studies identified during the primary search. This issue was investigated and addressed by modifying the search terms. To evaluate the review protocol, we got our SR protocol reviewed by an external expert in SRs in software engineering. The feedback helped us to improve the protocol.

Considering the descriptive nature of the literature to be retrieved, we also decided to perform a pre-review search to ascertain its general format and presentation and to inform the development of the data extraction and synthesis strategies. The pre-review search provided us with an initial indication of the scope of the literature available on the topic studied in this research.

2.2 Search strategy and data sources

The search strings used in this review were constructed using the following strategy:

- Derive main terms based on the research question and the topics being researched;
- Determine and include synonyms, related terms, and alternative spelling for major terms;
- Check the keywords in all relevant papers researchers already knew for example (Bachmann and Clements 2005; Svahnberg, van Gurp et al. 2005; Sinnema and Deelstra 2007), and initial searches on the relevant databases;
- Incorporate alternative spellings and synonyms using Boolean "or" and;
- Link main terms using Boolean "and";
- Pilot different combinations of the search terms.

Following this strategy, we constructed the search strings as bellow:

software AND (product line OR product lines OR product family OR product families) AND (variability OR variation OR variant)

The final search terms were constructed after a series of test executions and reviews. Due to the varying nature of the search features provided by the main digital sources of literature (such as IEEExplore, SpringerLink, and ACM Digital Library), it was not possible to use a single search string for all the digital sources. Hence, like other researchers (Kitchenham, Mendes et al. 2007), we also used different search strings for different sources with the exception of the ACM Digital Library where we had to construct three search strings in order to carry out search that can be considered an equivalent to other digital sources such as IEEExplore or Springer. Hence, whenever a database did not allow us to use complex search string involving several Boolean operators, we designed different search strings for each of those databases. This limitation of conducting SR is caused by the variations in the mechanisms of the search engines provided by the literature sources (Sjoberg, Dyba et al. 2007). We made every effort to ensure that the search strings used were logically and semantically equivalent if it was not possible to have syntactically identical search strings for all the searched databases. Three researchers were involved in designing and testing the search strings with different databases. All of them continuously discussed

and refined the search strings until they were fully satisfied with the capability of the used search strings. All the digital sources and their respective search strings can be found in (Chen, Babar et al. 2009).

We searched the primary studies in these digital sources (***1. IEEExplore; 2. ACM Digital library; 3. Citeseer library (Google); 4. ScienceDirect; 5. EI Compendex / Inspec; 6. SpringerLink; and 7. Web of Science***). As an indication of inclusiveness, the results were checked for three known relevant papers (i.e., (Svahnberg, van Gurp et al. 2005; Asikainen and Mannisto 2007; Sinnema and Deelstra 2007)). All three relevant papers were found in the search results. Apart from those electronic databases, we also manually checked two sources for candidate primary studies: ***(1. Proceedings of the SPLC conference series; and 2. SEI's technical reports on SPL)***. Note that SEI's serial of technical report is the main channel of grey literature in the research area under review. We did not restrict our search based on publication year. We performed the search in June 2007. That means the papers published after that date were not included in this study.

2.3 Study selection

Our search from all sources found 628 papers after removing the duplicates. These papers were downloaded into an Endnote library where all duplicates were removed by using the duplicate removal feature of Endnote. We performed a series of manual checks to ensure no duplicates remained in the library. We used a staged study selection process based on the inclusion and exclusion criteria according to which a paper was included if it:

1. introduces an approach to dealing with some aspect of VM in SPLE or;
2. reports an evaluation of an existing VM approach.

The paper was excluded if:

1. it does not deal with VM in SPLE.
2. it does not include an evaluation of a VM approach.
3. it is a short paper.

In the first stage, the 2 inclusion criteria and the first exclusion criteria were applied. After the first stage, 261 papers were selected for the next stage. In the second stage, the second exclusion criterion was applied, after which 146 papers qualified for the data extraction stage. Both stages were performed by a single researcher on the basis of title, abstract and conclusion of each paper. A second researcher randomly checked a small number of the selected primary studies during this process. To avoid missing relevant papers, we have decided to adopt a *greedy* strategy: include the paper, if the researcher has any doubt about excluding it.

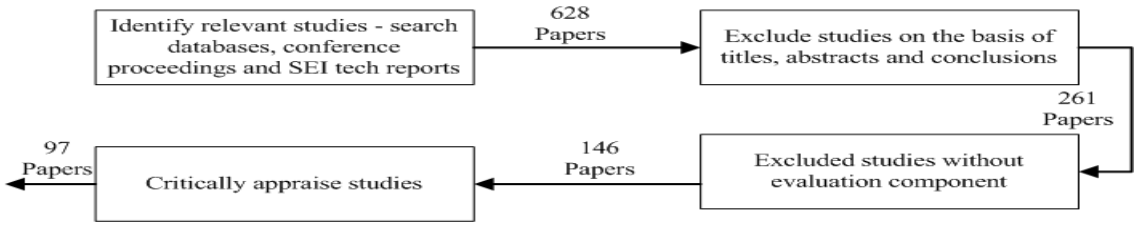

FIGURE 1: Study selection process

As shown in Figure 1, only 97 papers qualified for data extraction and appraisal purposes as we excluded 49 papers during the data extraction phase. These papers were excluded for two reasons: 1) the researchers included every paper for data extraction if they were in doubt about excluding a particular paper after reading the abstract and conclusion. However, a reading of the full paper enabled them to apply the selection criteria completely; 2) we decided to introduce the third exclusion criterion during the data extraction phase, when we found that the space limitation of short papers did not enable the authors to describe the full details of the presented approaches and their respective evaluations, if any. So finally 97 primary studies were included in this SR.

2.4 Quality assessment

We used two types of quality assessment in this study. The first type intends to assess the quality of the studies with respect to their ability and suitability to answer our research questions, and with respect to the impact on the drawn conclusions. The second type serves as an instrument to answer one of our main research questions. In a SR the main purpose of quality assessment is to assess the impact of the quality of the primary studies on the conclusions drawn from the SR. For example, if the quality of the primary studies is low, the conclusions based on these primary studies are not strong and reliable. Considering the objective of our research, our first kind of quality assessment treats each paper equally assuming that each of them was of sufficient quality as all but one of them were published in peer reviewed journals, conferences, and workshops. There was only one non-peer reviewed technical report published by SEI, which is considered highly credible institute for software engineering research. We will discuss the second kind of quality assessment used in this study later.

TABLE 1: Distribution of studies in different publication venues

ID	Pub. Channel	#	%	ID	Pub. Channel	#	%	ID	Pub. Channel	#	%	ID	Pub. Channel	#	%	ID	Pub. Channel	#	%
1	SPLC	11	11.3	**13**	ISESE	2	2.06	**25**	COMPSAC	1	1.03	**37**	ICSM	1	1.03	**49**	SEAS	1	1.03
2	SCP	6	6.19	**14**	JSS	2	2.06	**26**	EA	1	1.03	**38**	IEEE Computer	1	1.03	**50**	SEI-TR	1	1.03
3	PFE	5	5.15	**15**	NODe	2	2.06	**27**	ENTCS	1	1.03	**39**	IEEE Software	1	1.03	**51**	SoSyM	1	1.03
4	RE	5	5.15	**16**	REFSQ	2	2.06	**28**	EO	1	1.03	**40**	IRI	1	1.03	**52**	SPIP	1	1.03
5	ICSR	4	4.12	**17**	SPE	2	2.06	**29**	FACS	1	1.03	**41**	JASE	1	1.03	**53**	TOSEM	1	1.03
6	APSEC	3	3.09	**18**	AEI	1	1.03	**30**	FASE	1	1.03	**42**	MoDELS	1	1.03	**54**	TRI-Ada	1	1.03
7	ICSE	3	3.09	**19**	AuRE	1	1.03	**31**	GTTSE	1	1.03	**43**	MOMPES	1	1.03	**55**	TSE	1	1.03
8	SERA	3	3.09	**20**	CAiSE	1	1.03	**32**	HICSS	1	1.03	**44**	NJC	1	1.03	**56**	WCRE	1	1.03
9	ECBS	2	2.06	**21**	CAS	1	1.03	**33**	IASTEDSE	1	1.03	**45**	OOIS	1	1.03				
10	ESEC/FSE	2	2.06	**22**	CERE	1	1.03	**34**	ICCS	1	1.03	**46**	PROFES	1	1.03				
11	GPCE	2	2.06	**23**	CIT	1	1.03	**35**	ICESS	1	1.03	**47**	QSIC	1	1.03				
12	IEE Proceedings-Software	2	2.06	**24**	CN	1	1.03	**36**	ICFEM	1	1.03	**48**	SAICSIT	1	1.03				

Keys: Because of space limit, we omitted the full name of the publication channels here. Please refer to (Chen, Babar et al. 2009) for the full name of the publication channels.

TABLE 3: The kinds of VM approaches reported in the papers included in this SR

Nature of Solution	No. of Studies
Feature model	33
Using UML and its extensibility	25
Express variability as part of a technique that models the architecture of the system	8
Using natural language	6
Expressed variability as part of a technique that models the components of the system	5
Formal techniques based on mathematics	4
X-frames organized into a layered hierarchy	4
Domain specific language	3
Ontology based techniques	3
Solution from the perspective of Aspect-Orientation	2
Orthogonal Variability Management	2
Configuration management based Modelling	1
Using information visualization techniques	1

TABLE 4: The scheme for categorizing the evaluation approaches found in this SR

RA – Rigorous Analysis Rigorous derivation and proof, suited for formal model (Shaw 2003)
CS – Case Study An empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used (Yin 2002).
DC – Discussion Provided some qualitative, textual, opinion-oriented evaluation. E.g. compare and contrast, oral discussion of advantages and disadvantages (Carmen, Grigori et al. 2006)
EA – Example Application Authors describing an application and provide an example to assist in the description, but the example is "used to validate" or "evaluate" as far as the authors suggest (Shaw 2003).
EP – Experience The result has been used on real examples, but not in the form of case studies or controlled experiments, the evidence of its use is collected informally or formally (Shaw 2003).
FE – Field Experiment Controlled experiment performed in industry settings (Basili, Selby et al. 1986).
LH – Laboratory Experiment with Human Subjects Identification of precise relationships between variables in a designed controlled environment using human subjects and quantitative techniques (Harrison and Wells 2000)
LS – Laboratory Experiment with Software Subjects A laboratory experiment to compare the performance of newly proposed system with other existing systems (Glass, Vessey et al. 2002).
SI – Simulation Execution of a system with artificial data (Harrison and Wells 2000), using a model of the real word (Zelkowitz and Wallace 1998).

2.5 Data extraction and synthesis

During this stage, each of the 146 primary studies was fully read for data extraction purposes. For extracting the data, we had predefined a form consisting of a number of attributes, which were expected to be required in order to answer the research questions of this SR. Because of space limitation, this paper does not include the data extraction form, which consists of several attributes (i.e., reviewer's name, extraction date, title, authors, publication venue, publication year, publication source, research method used, variability management approach, evaluator, evaluation method, and industrial evaluation). Two researchers were involved in the data extraction phase. One researcher, called data extractor, was responsible for extracting and recording the data in an Excel Spreadsheet. Whenever the data extractor was in doubt, the paper was flagged to be checked by a second researcher. Where both researchers could not make a definitive decision, a third researcher was involved to clarify the doubts. It has been mentioned in Section 2.2 that during this stage 49 papers were excluded. So the data were extracted from 97 papers. Since the most of the selected primary studies were grounded in qualitative research, a meta-analytical approach was not suitable for synthesizing the data. We decided to manually review and link the extracted data in the Excel Spreadsheet. Then, we used descriptive statistics (e.g. sum, average) for analysing the data.

3. RESULTS

3.1 Overview of studies

3.1.1 Demographic data

With respect to the publication venues, Table 1 shows that the SPLC has the largest number of papers (11, 11.34%), followed by SCP journal (6, 6.19%), PFE (5, 5.15%), and RE (5, 5.15%) conferences. The primary studies have been appeared in 56 different publication venues. There are 39 venues with only one study published. These 39 venues also include TSE, TOSEM and IEEE Software. It was surprising to find that there was no study appeared in the Empirical Software Engineering Journal, which is a premier venue for publishing studies on empirical evaluation and assessment. These figures show that literature on VM is scattered in different publication venues. The premier events of the SPL community (i.e. SPLC, PFE) do not have clear dominance. It is quite interesting finding considering the fact that the most of the studies were published in conferences or workshops (75 of 97, 77.62%), while only 21 (21.65%) appeared in scientific journals and 1 in a technical report. Regarding the year of publication of the reviewed papers, Table 2 shows that there was no study on VM reported prior to 1990. There were only 4 studies before 2000. However, from 2000 onwards we found an increased number of studies with a peak in 2004.

TABLE 2: Publication chronology of the papers included in this SR.

Year	1990	1995	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	Total
Papers	1	1	1	1	6	3	8	7	23	15	15	16	97
%	1.03	1.03	1.03	1.03	6.19	3.09	8.25	7.22	23.71	15.46	15.46	16.49	100

3.1.2 Overview of VM approaches proposed

The study has identified 91 different approaches reported to deal with VM from different aspects and in different development phases. Considering the main objective of this SR, a detailed discussion on the classification of the VM approaches is not within the scope of this paper. However, we are performing an in-depth analysis of the presented approaches in our ongoing work. For this paper, we have categorized the identified approaches using a generic classification scheme based on the variability expression techniques used by each of the approaches. Table 3 shows the identified approaches and their relative frequencies (accumulated frequency is greater than 91 as some approaches can be placed into more than one category). Our results show that a large majority of the VM approaches are based on feature modelling and/or UML based techniques. There are also a small number of approaches based on some other mechanism of expressing variability such as natural language, mathematical notation and domain specific languages. It is important to note that there are very few approaches (4, 4.40%) based on mathematical techniques, which are usually considered hard candidate for empirical evaluation. That means a large majority of the VM approaches are quite amenable to empirical evaluation.

3.2 Evaluation methods used in primary studies

Software engineering community is emphasising the need for rigorously evaluating the approaches proposed for developing software (Shaw 2003; Carmen, Grigori et al. 2006). One of the main objectives of this SR was to identify the kinds of evaluation approaches being used by VM researchers. For classifying the evaluation

approaches reported in the primary studies reviewed, we did not find any suitable classification scheme available in the literature. For this study, we decided to develop a scheme for classifying the evaluation approaches used in the reviewed studies based on the work of Glass (Glass, Vessey et al. 2002), Shaw (Shaw 2003) and Carmen (Carmen, Grigori et al. 2006). The scheme for categorizing evaluation methods used in this SR is shown in Table 4.

Table 5 shows the kinds of evaluation approaches reported to support the VM approaches in the reviewed studies. It is evident that example application is the most frequently used means of evaluation followed by experience reports and case studies. Other evaluation approaches used are laboratory experiment with software subjects, laboratory experiment with human subjects, field experiment, simulation, rigorous analysis and discussion. It was observed that rigorous analysis has been applied when formal methods were used to solve the problem. It is also a notable finding that the majority the evaluation approaches used in the reviewed studies falls into the categories of example application, experience report, and discussion. They are not scientific approaches to rigorously evaluate a specific technology. We also found that the authors of a large majority the reviewed studies claimed to use case study methodology to evaluate their proposed approach. However, an analysis of those studies revealed that only a few of them could meet the requirements of case study research provided in (Yin 2002).

TABLE 5: Different evaluation approaches used.

Type of Evaluation	No. of Studies	Percent
Example Application	57	58.76
Experience Report	17	17.53
Case Study	13	13.4
Discussion	4	4.12
Laboratory Experiment with Human Subjects	1	1.03
Simulation	1	1.03
Laboratory Experiment with Software Subjects	2	2.06
Field Experiment	1	1.03
Rigorous Analysis	1	1.03
SUM	97	100

We also observed that many authors claimed to have their approaches evaluated using industrial case studies, however, they provided a paragraph or less on the evaluation part of the reported approach. An author claimed to use industrial experiment, which was described in one sentence. These findings reveal that a large majority of the VM approaches awaits rigorous empirical evaluation, which is vital for successful technology transfer (Dyba, Kitchenham et al. 2005). To investigate whether there are trends in the attempts at empirically evaluate proposed approaches, we analyzed the temporal distribution of the evaluation approaches used in the reviewed studies; however, we did not find a clear trend of improvement. It was also found that a large number of recently proposed VM approaches do not provide any rigorous evaluation of their utility, which is very disappointing. Another significant finding was the absence of any replication studies. 87 (96%) of the presented approaches were evaluated in only one study. There was only one study (Christian and Ronny 2006) that can be considered as an independent evaluation. Rest of the approaches have been evaluated by their developers.

TABLE 6: Industrial evaluation of the VM approaches reported in the papers included in this SR

Industrial evaluation/trial	No. of Approaches	Percent
Not evaluated in industrial settings	65	71.43
Tried in industrial settings	26	28.57
SUM	91	100

3.3 Evaluation performed in an industrial setting

The purpose of software engineering research is to provide practitioners with solutions to real problems (Fenton, Pfleeger et al. 1994). Our SR was also aimed at finding out the number of approaches evaluated in an industrial context. Table 6 presents the results from this analysis. It is event that the large majority of the reviewed approaches have never been evaluated in an industrial setting. Among those approaches that have been tried in industrial settings, more than half of them were reported as experience reports. We also found that it was not clear how many of those approaches were adopted in industry based on those trials. However, we observed from the available data that only a small number of the approaches claimed to be evaluated in industry was accepted. One

common characteristic of these approaches was that they were developed in close collaboration with industrial partners. Hence, this finding supports the claim that industrial acceptance of a technology is a strong indicator of the success of the research output.

3.4 Quality of evaluation

We also intended to assess the quality of the evaluation studies reported to support the VM approaches. For this objective, we decided to perform an in-depth analysis of the selected primary studies using the quality assessment criteria adopted from another SR (Dyba and Dingsoyr 2008; Dybå and Dingsøyr 2008). We only made a few minor changes to customize the detailed sub-criteria presented in Appendix B of (Dyba and Dingsoyr 2008) for our study. A summary of the quality assessment criteria is presented in Table 7.

TABLE 7: Quality criteria (adapted from (Dyba and Dingsoyr 2008))

1. Is the paper based on research (or is it merely a ‘‘lessons learned” report based on expert opinion)?
2. Is there a clear statement of the aims of the research?
3. Is there an adequate description of the context in which the research was carried out?
4. Was the research design appropriate to address the aims of the research?
5. Was the recruitment strategy appropriate to the aims of the research?
6. Was there a control group with which to compare treatments?
7. Was the data collected in a way that addressed the research issue?
8. Was the data analysis sufficiently rigorous?
9. Has the relationship between researcher and participants been considered to an adequate degree?
10. Is there a clear statement of findings?
11. Is the study of value for research or practice?

Because many studies are not suitable to be rated as either “OK” or “No”, we decided to use a ternary (“OK”, “Partial” or “No”) scale to grade the studies on each of the criterion element in the quality assessment criteria. To quantify the result, we assigned these values: 1 to OK, 0.5 to Partial and 0 to No.

According to the quality criteria, if the answer to the first criterion was “No”, the most of the remaining criteria are not applicable. Hence, we decided not to apply the whole quality assessment criteria to the studies using Example Application (EA), Experience Reports (ER) and Discussion (DC) as the means of evaluation of the proposed approaches. The criteria seem not applicable to studies using rigorous analysis either. This decision left us with only 18 studies to be assessed against the quality criteria. Because we only applied the 11 elements of the criteria on the selected papers, all studies were rated as OK on the first criterion. However, there were only 7 studies with a clear statement of the aims of the reported research; others were rated as Partial. All studies had some form of description of the context in which the research was carried out, but the context description of one study was not clear enough to gain a reasonable understanding. Only 3 studies provided both justification and description of the research design, all other studies had some sort of description of the research design. There were only 6 studies that reported an appropriate recruitment strategy, rest of the 11 studies went directly to the description of the participants or cases without explaining how and why they were identified and selected. Only 3 studies included one or more control groups. Only 7 and 5 studies, respectively, provided the description of their data collection protocols and data analysis procedures; others studies mentioned the data collection and analysis protocols or procedures but did not provide any explanation. None of the reported studies mentioned any possibility of researchers’ bias. Only 5 studies described both the findings and discussed validity threats to the reported research, other 13 studies did not discuss any validity threats at all.

TABLE 8: Quality assessment of 18 empirical studies

	1	2	3	4	5	6	7	8	9	10	11
	Research	Aim	Cxt	R.design	Sampling	Ctrl Grp	Data coll	Data anal	Reflex.	Findings	Value
OK	18	7	17	3	6	3	7	5	0	5	18
Partial	0	11	1	15	11	0	8	13	0	13	0
No	0	0	0	0	1	15	3	0	18	0	0
AVG	1	0.69	0.97	0.58	0.64	0.17	0.61	0.64	0	0.64	1

Table 8 shows the average score assigned for each criterion. It is evident that there is a significant gap regarding the recognition of researcher bias (scored 0) and control group to compare treatments (scored 0.17) respectively. Except the first, last criterion and the criterion regarding description of the context, there is a huge gap (scored around 0.60) that needs to be filled in order to improve the quality of the evaluation of the VM approaches.

4. DISCUSSION AND LIMITATIONS

The findings of this SR have revealed that VM research has produced quite diverse approaches (i.e. 91 different approaches), however, a large majority of them await rigorous empirical evaluation. Hence, a major concern that this SR has highlighted is a general lack of rigorous evaluation of VM approaches. The majority of the reviewed approaches did not appear to be sufficiently evaluated as they do not provide any evidential support for the claimed utility of the proposed approach. The results also show the use of a diverse set of evaluation approaches used to provide the evidence for supporting the presented VM approaches. The evaluation approaches used by a large majority of the researchers can be categorized as example application.

The prevalence (80.41%) of the scientifically not rigorous evaluation approaches (such as example application, experience report and discussion) indicates a general lack of robust assessment of the large majority of the VM approaches. It is particularly worth mentioning that some of the description of the evaluation was just one line of statement without any detailed information provided. Apart from a general lack of rigorous evaluation, a detailed investigation of the studies employing empirical research methods revealed huge quality deficiencies in the majority of the reported studies on most (8 of 11, 72.73%) of the elements of the quality assessment criteria used. However, it should also be noted that some of these approaches were in their early stages of the maturity. Some authors also reported that they intended to rigorously evaluate the proposed approaches in future. So it may be worthwhile to track those papers in future to verify whether or not such claims are followed.

The results also reveal a lack of number of studies carried out to evaluate a particular approach as 95.60% of the proposed approaches were evaluated by only one study. We did not find any replicated studies at all. The results also show that almost all of the studies, except one (Christian and Ronny 2006), were conducted by the researchers who proposed the approaches. Unarguably, this may bring in some bias and subjectivity to the reported effectiveness of the proposed approaches. Moreover, without independent evaluation and replication, it is unlikely to develop any useful theoretical foundations for a phenomenon or provide solid and reliable evidence to support a particular technology. Unfortunately, most of the evaluation studies of VM approaches reviewed in this SR do not show any depth of enquiry or a strong evidence to support the claims made in support of the proposed approaches.

Another important issue that this SR has also discovered is the lack of industrial evaluation of the proposed approaches. Hence, like other disciplines of software engineering (e.g., requirements and software architecture), evaluation of a large majority of the VM approaches remains an open research issue, which needs much more efforts from researchers and practitioners than it has received over the last 18 years of VM research and practice.

With respect to the publication channels, the primary studies are quite scattered, and the dominance of some well-known publication channels in SPLE is not that clear. This result is somehow out of our expectation. The implication to researchers was that limiting the scope of search of primary studies to a short list of well-known publication channels may miss a large number of primary studies, which can only be easily detected with a comprehensive SR like ours.

When excluding the papers without an evaluation component during the paper selection stage, we also paid attention to the potential impact of the papers based on the number of citations reported on Google Scholar. Interestingly, we noticed that some of the papers that proposed an approach but did not have an evaluation component were excluded from our SR, but they were highly cited. We analyzed the number of citations of those papers (Bachmann and Bass 2001; Batory, Lopez-Herrejon et al. 2002; Muthig and Atkinson 2002; W. Krueger 2002; Muccini and Van Der Hoek 2003; Bachmann, Goedicke et al. 2004; Czarnecki, Helsen et al. 2004). All of them got more than 20 citations. The average number of citations is 40.86, which is a very high value in this research community and apparently will overweight the average number of citation of those papers with an evaluation component included in the reported SR. One may interpret this finding as a sign of a general lack of appreciation of the importance of empirical studies in this community.

We also found that the papers in the area of variability management of software product lines rarely reference empirical research methodologies published by researchers in empirical software engineering community. We suspect that the researchers in the VM research community are not aware of the available empirical evaluation methods reported by the empirical software engineering community, or they don't recognize the importance of rigorous empirical evaluations.

In a nutshell, the findings from this SR reveals that the status of evaluation of VM approaches is not satisfactory rather poor in certain aspects. Not only the available evidence is sparse, but the quality of the reported evaluations is very low. Hence, any estimate of effect that is based on evidence of VM approaches from current research can hardly be considered reliable. Considering the fact that a large majority of the approaches fall into the category of informal methods, empirical evaluation and assessment are expected to play a vital role. While the findings of this SR has several implications for VM researchers and practitioners, it also identifies the areas where empirical software engineering community can be encouraged to work with SPLE community to improve the state of the practice of rigorously evaluating research outcomes.

Limitations: The findings of this SR may have been affected by certain limitations such as bias in selection of publications, inaccuracy in data extraction and reliability of classifying the evaluation approaches reported and quality assessment. We tried our best to search all papers that have been published in the literature on VM in SPLE. However, it is possible that we may have not found those studies whose authors might have used other terms for the same thing in the early stages of research on VM or because of the reasons reported by other researchers for their SRs (Dyba and Dingsoyr 2008; Kampenes, Dybå et al. 2009). Our SR may also have missed those VM approaches that have been commercialized but have not been reported in literature with an evaluation component. Since we mainly relied on search engines to retrieve the primary studies, the quality of search engines could have influenced the completeness of the identified primary studies. We also found that many articles lacked sufficient details about the design and execution of the reported studies. That was why sometimes we had to infer certain pieces of the required information. There is therefore a possibility that the extraction process may have resulted in some inaccuracies. The classification process involved subjective decisions by the researchers. To minimize these limitations, a third researcher checked the papers about which the primary researchers were in doubt or felt uncertain. All discrepancies were resolved through reviews of the papers and discussions. The subjectivity involved in assessing the quality of the reported studies is another limitation that should be taken into account while interpreting the results of this SR.

5. CONCLUSION AND FUTURE WORK

To review the status of evaluation of VM approaches in software product line engineering, we have carried out a systematic review of the studies of VM in SPLE reported in any publication venue and published before the middle of 2007. The results presented in the previous sections provide interesting insights into the current status of VM research with respect to evaluation of the proposed approaches. We have observed a wide diversity in the proposed solutions to manage variability, lack of rigorous and high quality evaluation and industrial trials of the proposed approaches. We sketched the status of evaluation of VM approaches and identified gaps and areas that needs to be filled by future research.

We believe that this paper provides an important contribution to both practitioners and researchers as it can provide them with useful information about different aspects of the VM research outputs. Particularly, the low quality of evaluation studies revealed by this SR can help practitioners to correct misunderstanding and over interpretation of the reported evidence in the literature; in the meanwhile, the revelation of paucity of empirical evidence and the scattered distribution of papers over a large number of publication venues can be useful information for researchers who are going to do literature review in this area. The results also highlight the areas, which need immediate attention by researchers and practitioners as more active collaboration between these two communities is expected to result in VM technologies, which would have higher potential of industrial adoption. Moreover, the findings of this SR should also be of interest to the Empirical Software Engineering (ESE) community as there is a vital need for conducting high quality empirical studies on VM approaches reported in the literature and the ESE community is well placed to make significant contributions in this respect by performing independent evaluation of the available VM approaches.

We intend to extract more data and perform further analysis to provide more in-depth insights to understand the specific areas (such as requirements, architecture, design, realisation and testing) for which different kinds of VM approaches have been developed and how these approaches aim to meet industry requirements. From the evidence synthesis point of view, although the evidence is sparse and the quality of the studies is low, an in depth synthesis and analysis of the reported evidence is still possible to yield interesting insights into the consistency of the provided evidence, the reported benefits and limitations of VM approaches and the factors that can influence the effectiveness of those approaches. Our future work will look into these aspects of the VM literature.

6. ACKNOWLEDGEMENTS

The work was supported, in part, by Science Foundation Ireland grant 03/CE2/I303_1. We would also like to thank Prof. Barbara A. Kitchenham for her help in reviewing our research protocol.

REFERENCES

Asikainen, T. and T. Mannisto (2007). "Kumbang: A domain ontology for modelling variability in software product families." Advanced Engineering Informatics **21**(1): 23-40.

Bachmann, F. and L. Bass (2001). "Managing variability in software architectures." SIGSOFT Softw. Eng. Notes **26**(3): 126-132.

Bachmann, F. and P. Clements (2005). Variability in Software Product Lines, Software Engineering Institute, Pittsburgh, USA.

Bachmann, F., M. Goedicke, et al. (2004). "A Meta-model for Representing Variability in Product Family Development." Lecture Notes in Computer Science **3014/2004**: 66-80.

Basili, V. R., R. W. Selby, et al. (1986). "Experimentation in software engineering." IEEE Trans. Softw. Eng. **12**(7): 733-743.

Batory, D., R. E. Lopez-Herrejon, et al. (2002). Generating product-lines of product-families. Automated Software Engineering, 2002. Proceedings. ASE 2002. 17th IEEE International Conference on.

Bosch, J. (2000). Design & Use of Software Architectures: Adopting and evolving a product-line approach, Addison-Wesley.

Bosch, J., G. Florijn, et al. (2002). "Variability Issues in Software Product Lines." Lecture Notes in Computer Science **2290/2002**: 303-338.

Carmen, Z., M. Grigori, et al. (2006). On the success of empirical studies in the international conference on software engineering. Proceedings of the 28th international conference on Software engineering. Shanghai, China, ACM.

Chen, L., M. A. Babar, et al. (2009). A Survey of Evaluation of Variability Management Approaches, Lero Technical Report, Limerick, Ireland (http://www.staff.ul.ie/alibabar/Review-VM.pdf).

Christian, D. and K. Ronny (2006). Testing and inspecting reusable product line components: first empirical results. Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering. Rio de Janeiro, Brazil, ACM.

Clements, P. and L. Northrop (2002). Software Product Lines: Practices and Patterns, Addison-Wesley.

Czarnecki, K., S. Helsen, et al. (2004). Staged configuration using feature models. Software Product Lines, Proceedings. Berlin, Springer-Verlag Berlin. **3154:** 266-283.

Dyba, T. and T. Dingsoyr (2008). "Empirical studies of agile software development: A systematic review." Information and Software Technology **50**(9-10): 833-859.

Dybå, T. and T. Dingsøyr (2008). Strength of evidence in systematic reviews in software engineering. Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement. Kaiserslautern, Germany, ACM.

Dyba, T., B. Kitchenham, et al. (2005). "Evidence-Based Software Engineering for Practitioners." IEEE Software **22**(1): 58-65.

Fenton, N., S. L. Pfleeger, et al. (1994). "Science and substance: a challenge to software engineers." Software, IEEE **11**(4): 86-95.

Glass, R. L., I. Vessey, et al. (2002). "Research in software engineering: an analysis of the literature." Information and Software Technology **44**(8): 491-506.

Harrison, R. and M. Wells (2000). A Meta-analysis of Multidisciplinary Research. The Conference on Empirical Assessment in Software Engineering (EASE)**:** 1-15.

Kampenes, V. B., T. Dybå, et al. (2009). "A systematic review of quasi-experiments in software engineering." Information and Software Technology **51**(1): 71-82.

Kitchenham, B. and S. Charters (2007). Guidelines for Performing Systematic Literature Reviews in Software Engineering, Keele University, UK.

Kitchenham, B., T. Dyba, et al. (2004). Evidence-Based Software Engineering. Proceedings of the 26th International Conference on Software Engineering, IEEE Computer Society.

Kitchenham, B., E. Mendes, et al. (2007). "Cross versus within-Company Cost Estimation Studies: A Systematic Review." IEEE Transactions on Software Engineering **33**(5): 316-329.

Muccini, H. and A. Van Der Hoek (2003). "Towards testing product line architectures." Electronic Notes in Theoretical Computer Science **82**(6): 109-119.

Muthig, D. and C. Atkinson (2002). Model-Driven Product Line Architectures. Proceedings of the Second International Conference on Software Product Lines, Springer-Verlag.

Schmid, K. and I. John (2004). "A customizable approach to full lifecycle variability management." Science of Computer Programming **53**(3): 259-284.

Shaw, M. (2003). Writing good software engineering research papers. Software Engineering, 2003. Proceedings. 25th International Conference on.

Sinnema, M. and S. Deelstra (2007). "Classifying variability modeling techniques." Information and Software Technology **49**(7): 717-739.

Sjoberg, D. I. K., T. Dyba, et al. (2007). The Future of Empirical Methods in Software Engineering Research. Future of Software Engineering, 2007. FOSE '07.

Svahnberg, M., J. van Gurp, et al. (2005). "A taxonomy of variability realization techniques." Software-Practice & Experience **35**(8): 705-754.

W. Krueger, C. (2002). Variation Management for Software Production Lines. Proceedings of the Second International Conference on Software Product Lines, Springer-Verlag.

Yin, R. K. (2002). Case Study Research: Design and Methods. Thousand Oaks, CA, Sage Publications, Inc.

Zelkowitz, M. V. and D. R. Wallace (1998). "Experimental models for validating technology." Computer **31**(5): 23-31.