On the Reliability of Mapping Studies in Software Engineering

Claes Wohlin[1], Per Runeson[2], Paulo Anselmo da Mota Silveira Neto[3], Emelie Engström[2], Ivan do Carmo Machado[4] and Eduardo Santana de Almeida[3, 4]

[1]School of Computing,
Blekinge Institute of Technology
371 79 Karlskrona
Sweden

[2]Department of Computer Science,
Lund University
221 00 Lund
Sweden

[3]RiSE - Reuse in Software Engineering
Recife, PE
Brazil

[4]Computer Science Department
Federal University of Bahia
Salvador, BA
Brazil

Abstract

Background: Systematic literature reviews and systematic mapping studies are becoming increasingly common in software engineering, and hence it becomes even more important to better understand the reliability of such studies.
Objective: This paper presents a study of two systematic mapping studies to evaluate the reliability of mapping studies and point out some challenges related to this type of study in software engineering.
Method: The research is based on an in-depth case study of two published mapping studies on software product line testing.
Results: We found that despite the fact that the two studies are addressing the same topic, there are quite a number of differences when it comes to papers included and in terms of classification of the papers included in the two mapping studies.
Conclusions: From this we conclude that although mapping studies are important, their reliability cannot simply be taken for granted. Based on the findings we also provide four conjectures that further research has to address to make secondary studies (systematic mapping studies and systematic literature reviews) even more valuable to both researchers and practitioners.

Keywords

Systematic mapping study, systematic map, software product lines, systematic reviews, systematic literature review, review of reviews, meta-study, software testing

1. Introduction

Research is published to enable others to build on the conducted research. There are two major approaches to building on the research of others, either continuing the type of research presented or using the results presented to synthesize the knowledge of research in an area by systematically processing and structuring the research. The former means adding more primary studies while the latter involves performing secondary studies, i.e. studies where the results from several primary studies are collated. Secondary studies are a standard approach in many research disciplines. The advent of a systematic approach to secondary studies in software engineering was in 2004 [Kitchenham04] when Evidenced-based Software Engineering (EBSE) was introduced as a concept. It was later adapted to software practitioners [Dybå05]. The approach taken on EBSE was inspired by the practice in evidence-based medicine, although secondary studies have been used in Information Systems [Webster02], which is a discipline closer to software engineering than evidence-based medicine. Thus, the objective of a secondary study is to provide both researchers and practitioners with an overview of an area, and to identify gaps in the literature. Researchers may use secondary studies to direct their work to research gaps; practitioners may use secondary studies to understand the effectiveness and efficiency of, for example, a specific method or technology. Secondary studies are needed to make the best possible decisions related to future research and practice. This implies that the secondary studies must be perceived as reliable and trustworthy.

In this context, it is worth noting that a well-defined approach to literature searches should be viewed as the way to conduct literature reviews, whether being for a secondary study or being a literature review for any research paper. The introduction of worldwide accessible research databases and search engines has changed the expectations of literature studies independently of the objective of the search as such.

Since the introduction of EBSE and the development of guidelines for systematic reviews [Kitchenham07], secondary studies in software engineering have increased over the years as illustrated by several tertiary studies [Kitchenham09, Kitchenham10, da Silva11]. Secondary studies may be divided into systematic literature reviews and systematic mapping studies (or scoping studies). A systematic literature review is expected to provide a synthesis of the knowledge in an area (often related to a specific research question), while a mapping study primarily maps out the research by different classifications of the primary studies included in the secondary studies. Thus, the main difference is in scope and in the analysis procedures. The scope is often broader in a mapping study than in a systematic literature review, and a deeper synthesis is not expected in a systematic mapping study. However, we would argue that the procedures for selecting primary studies to include in the secondary study should be the same for systematic literature reviews and systematic mapping studies, although Kitchenham et al. on the one hand expect the search strategy to be “less stringent” for mapping studies and on the other hand stress the need for high quality in completeness and rigour for them to form a good basis for further research [Kitchenham11b]. When Kitchenham et al. refer to less stringent, the examples given relate to limitations in terms of the sources used for relevant papers. Basically, a systematic map may be viewed as a subset of a systematic literature review, i.e. a study without the synthesis step and often addressing an area rather than a more specific research question. Kitchenham et al. [Kitchenham11b] elaborate further on the differences between reviews and maps.

With the increase in the number of published secondary studies, it is obvious that it becomes important to also understand their reliability. Reliability in this context relates to whether two studies on the same topic would draw the same conclusions about a research area, where the overlap in papers identified is one factor. To enable such an evaluation, more than one study is needed on a topic. MacDonell et al. designed a study to evaluate the reliability of systematic literature reviews [MacDonell10] with two teams conducting the same systematic review on estimation for cross-company and within company models. Experts in the area conducted the review and it is a quite narrow topic. The two teams came to similar findings, and hence they concluded that in their case the systematic literature review was reliable. If it had not worked in this case, it would be hard to believe that systematic literature reviews would be reliable in a broader context. The challenge is increased in terms of reliability when we have two studies without a common research question, although addressing the same topic. Here, we present a study comparing two independent mapping studies. The study is based on two systematic mapping studies on software product line testing [Engström11, Neto11], which were conducted by two independent research teams. The researchers were unaware that another team was conducting a similar mapping study. Thus, the study here is opportunistic in the sense that the opportunity to study the reliability came from the availability of two systematic maps and not from actually designing a study with two teams. The first author of this paper initiated the reliability study and invited the other authors since more in-depth knowledge of the two systematic mapping studies was needed than provided in the published maps. Thus, the study presented may be positioned as a participant-observer case study in a similar way as done by, for example, Kitchenham et al. [Kitchenham12a].

In the study by Kitchenham et al. [Kitchenham12a], the authors conducted a mapping study of empirical studies in unit testing and regression testing. The authors compared the papers identified with those from an expert literature review and six mapping studies or literature reviews on partially overlapping topics. None of the studies addressed exactly the same area as the mapping study conducted by the authors. Thus, to the best of our knowledge, this paper contributes the first reliability study of two independent systematic mapping studies in software engineering focused on the exact same topic. It provides an in-depth analysis of two systematic mapping studies in the area of software product line testing. It compares the goals for the studies, the papers included in the two mapping studies and in particular the classification of the papers found in both mapping studies. Based on the experiences from the analysis, the paper provides a number of reflections and concludes by presenting a set of conjectures regarding secondary studies in software engineering. The conjectures provide a basis for further development of the methods for conducting secondary studies.

In summary, the main focus of this paper is to evaluate the reliability of mapping studies by comparing the outcome of two independent systematic mapping studies on the same topic. In particular, the objective is to compare the two mapping studies both with regard to the papers identified and the classification of the identified papers. The following definitions are used for systematic mapping studies, systematic literature review and reliability respectively in the paper. The definition of a systematic literature review is included to contrast it with a systematic mapping study.

Definitions:
"Systematic mapping study (also referred to as a scoping study): A broad review of primary studies in a specific topic area that aims to identify what evidence is available on the topic." [Kitchenham07]

"Systematic literature review: (also referred to as a systematic review). A form of secondary study that uses a well-defined methodology to identify, analyse and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable." [Kitchenham07]

"Reliability: Demonstrating that the operations of a study – such as the data collection procedures – can be repeated, with the same results." [Yin09]

Given the definitions of systematic mapping studies and reliability, we conclude that two mapping studies on the same topic ought to produce a similar overview of a topic for it to be possible to claim that mapping studies are indeed reliable, including that the findings are repeatable and consistent. We study this issue by comparing the goals of the studies, listing all the primary studies of the two mapping studies, analysing differences and similarities between the sets of primary studies, as well as how they are classified in the respective mapping studies. Our findings lead us to identify conjectures that may lead to more reliable secondary studies in the future.

The remainder of the paper is outlined as follows. Section 2 presents related work. In Section 3, the research method and research questions are presented. The two mapping studies forming the input to the analysis are introduced in Section 4. The results of the analysis are presented in Section 5, which is followed by a discussion based on the findings in Section 6. Section 7 presents the conclusions and the conjectures based on the study. Finally, all papers included in the systematic maps are listed in a structured way in Appendix A and in Appendix B the classification of the research type of the papers found by both studies is provided.

2. Related work

Systematic literature reviews have become established in the field of medicine as a way of synthesizing evidence and then ultimately allowing researchers to come to a joint understanding of the status of a research area. In clinical medicine, the studies are often randomized controlled trials, which then makes synthesis easier (although not necessarily easy) than if aiming to synthesize studies including more qualitative data and with more variation in the context [Cruzes11]. Case studies are one such example. Inspired by medicine, the concept of evidence-based software engineering was coined by Kitchenham et al. [Kitchenham04]. In parallel, similar ideas have been brought into information systems research, e.g. by Webster and Watson [Webster02].

Evidence-based software engineering may be viewed as an evolution of research discussing the need to synthesize research results as discussed in the late 1990s [Pickard98, Miller99, Hayes99]. Pickard et al. discuss combining research results, Miller addresses the issue of combining research results through meta-analysis and Hayes uses the concept of synthesis of research results. They all have in common that they stress the need for a systematic approach to not only conducting individual research studies, but also to building knowledge from combining findings from different studies on a topic. One such early example is the work by Basili et al. [Basili99], where the authors look into combining the research and hence knowledge we have regarding research on software inspections.

Based on the original EBSE ideas [Kitchenham04], the research related to systematic literature reviews, and more generally secondary studies in software engineering, has subsequently evolved. The ideas have been formulated from a practitioner's point of view [Dybå05] and guidelines for

conducting systematic literature have been developed [Kitchenham07]. Furthermore, the concept of using systematic mapping studies as a complement to systematic literature reviews has been proposed [Petersen08]. Kitchenham et al. [Kitchenham11b] discuss the use of systematic mapping studies as a starting point for further research. We use secondary studies as a collective term for systematic literature reviews and systematic mapping studies.

The research process for systematic mapping studies has five main steps [Petersen08]: 1. Defining research questions. 2. Conducting the search for primary studies. 3. Screening papers based on inclusion/exclusion criteria. 4. Classifying the papers. 5. Data extraction and aggregation. The research questions for mapping studies are generic and related to research trends, typically of the form "which researchers, how much activity, what type of studies, etc." [Kitchenham11b].

As a result of the evolution and introduction of secondary studies in software engineering, the number of secondary studies published has grown in the last five years [Kitchenham09, Kitchenham10, da Silva11]. A consequence of having more secondary studies published is that it is possible to investigate secondary studies as such [Kitchenham09] and also study specific aspects of secondary studies such as repeatability, in particular in relation to novice researchers [Kitchenham11a], and the reliability of, for example, systematic literature reviews [MacDonell10]. In the latter study, the authors divided themselves into two teams to conduct two independent systematic literature reviews from a joint research question. The study resulted in a large overlap between papers found in the two systematic literature reviews. In Kitchenham et al. [Kitchenham12a], the authors conducted a mapping study of empirical studies related to unit testing and regression testing, and the papers found were compared with an expert literature review and six other secondary studies with overlapping topics. When comparing the papers found, they conclude that they missed at least 10% of the relevant papers in relation to the other studies and in some cases a large number of papers. It is also concluded that the other secondary studies also missed papers.

Overlap between secondary studies has been studied in other fields too. Some examples are reported here. Woodman et al. conducted a study of eight reviews reporting evidence in the area of physical activity interventions [Woodman12]. The overlap of primary studies was lower than anticipated by the authors. Most primary studies were only found in one of the reviews. Ekeland et al. did a review of systematic reviews regarding the effectiveness of telemedicine [Ekeland10]. They found that the reviews came to different conclusions about the effectiveness, which may be due to differences in judging, for example, costs and patient benefits. However, they also concluded that several reviews studied similar or overlapping topics, and they observed that the reviews at least partially included the same papers. Taken together, the reliability of secondary studies may be a challenge, and hence it is important to address this further within software engineering.

The evolution of secondary studies in software engineering has now reached a state where it is possible to improve the way secondary studies are conducted, as illustrated by investigations of the search strategy [Dieste09, Skoglund09, Zhang11, Jalali12]. Another example is that researchers have questioned the way systematic literature searches are conducted [Boell11]. They argue that the actual search is not the most important step; it is the reading and understanding of an area that is the key to a good literature review. The latter is aligned with the findings that novice researchers may not be very well equipped to conduct secondary studies [Kitchenham11a].

Despite the increased focus on conducting systematic literature studies in software engineering, there is still too little attention on conducting research synthesis. This needs to change and synthesis needs to be an integral part of systematic reviews to increase their significance and usefulness for research and practice [Cruzes11]. An important step in this direction is to investigate and understand how secondary studies should be interpreted and hence in particular the reliability of them. Thus, based on the publication of two independent mapping studies [Engström11, Neto11], this paper provides some new insights on the reliability of systematic mapping studies in software engineering. The research complements the findings from earlier studies [MacDonell10, Kitchenham11a, Kitchenham12a].

It should be acknowledged that these two mapping studies are not the first studies focusing on systematically mapping or reviewing the state of the art in testing of software product lines. Lamancha et al. [Lamancha09] presented a systematic review[1], which resulted in 23 primary studies being identified for inclusion. The primary studies were classified into seven areas: Unit Testing, Integration Testing, Functional Testing, SPL Architecture Testing, Embedded Systems Testing, Testing Process and Testing Effort in SPL. It then discusses variability points and testing technique used. However, the study by Lamancha et al. cannot be compared to the two mapping studies, since the study was available to both research groups when performing their systematic mappings. Thus, it does not qualify as an independent secondary study in this analysis.

3. Research method

Recognising that two systematic mappings on the same topic had been accepted for publication, it was decided that it was a unique opportunity to analyse them together. The actual research method may be classified as being a participant-observer case study with two cases. Others also use this type of case study methodology when doing this type of analysis as exemplified by [Kitchenham11b, Kitchenham12a]. The first author has acted as an observer of the research throughout the work (not involved in any of the two mapping studies), and the other authors have been participants in their own systematic mapping study, and observers of the other mapping study. The two cases in the case study are the two independent systematic mapping studies. The case study includes two data collection methods, i.e. archival analysis of the two systematic mapping studies and interviews via email with the authors of the two mapping studies.

The first author of this paper has been the main driver of the comparison of the two mapping studies to ensure as much impartiality in relation to the original mapping studies as possible. Initially, the intention was to base the study solely on the published papers to ensure objectivity. However, it was concluded that some additional input was needed and that it would be beneficial if the authors of the mapping studies were involved to help in ensuring that the papers were interpreted correctly. Thus, one of the more senior authors of each of the mapping studies was contacted (Runeson and de Almeida), and that person has then also received support from some of the other authors of the mapping studies.

[1] According to the definitions in this paper, it is a mapping study.

3.1 Research question

The main research question is: Are similar results obtained when independent researchers systematically create a systematic map? This research question is a variant of the research question posed by MacDonell et al. [MacDonell10]: "How reliable are systematic reviews?"

This question is further broken down into:

RQ1: Which steps of the mapping studies are shown to be reliable or unreliable and why?

RQ2: How are the general conclusions about the topic affected by differences in the two independent studies?

RQ3: Based on the analysis, what recommendations can be made to help improve reliability of secondary studies such as systematic reviews and systematic mapping studies?

The word "reliable" is used here to mean both repeatable and consistent. A secondary study should be repeatable in the sense that it should be possible to redo a study and understand the reasoning of the authors. This is not the same as being consistent. Consistency is in this context a matter of conducting an independent new systematic study and obtaining similar results.

3.2 Research process

The research process was as follows:

1. The first author extracted as much information as possible from the two systematic maps, including listings of papers and a summary of the process used in the two mapping studies.
2. The contact authors were asked to explain whether papers found in the other mapping study were never found or excluded in the evaluation process. Furthermore, they were asked to check the summary of the procedure used in their respective mapping study.
3. The contact authors were asked to explain the process when classifying the research papers into research types, see Section 5.2. A naming issue regarding one research type was resolved.
4. Based on the input, the first author summarized the material in the paper and it was sent for review to the contact authors. Some clarifications were introduced based on the feedback. Based on the agreement of the interpretation of the two mapping studies, the first author continued the analysis.
5. The paper was iterated several times for revision and commenting.

To maintain objectivity, the authors of the original mapping studies were initially only asked questions; they were then provided with the summaries of the mapping studies to check them for correctness, and finally they were also provided with the actual results. The whole research process was organized via email.

4. The two mapping studies

4.1 Introduction

This comparative study is based upon two recent mapping studies on software product line testing. Both articles were published in Information and Software Technology. The two studies were conducted in parallel by two independent research teams: one team was located in Sweden and one

team in Brazil/USA. Both studies follow the normal patterns when it comes to reporting, i.e. the mapping studies report classifications on the research in the area, but they do not provide any real synthesis of the primary studies. Thus, it is impossible to go beyond comparing the classifications here. Once a synthesis is reported then it is commonly denoted as a systematic literature review instead of a systematic mapping study.

It is particularly interesting to compare the two studies since they use the same classification scheme for describing the research type. Wieringa et al. introduced this classification scheme for papers in requirements engineering [Wieringa06], but the scheme is not very specific to that field and may be used more generally. In summary the six research types are defined as follows:

- *Evaluation Research:* Techniques, methods, tools or other solutions are implemented and evaluated in practice, and the outcomes are investigated.
- *Validation Research:* A novel solution is developed and evaluated in a laboratory setting.
- *Solution Proposal:* A solution for a research problem is proposed, and the benefits are discussed, but not evaluated.
- *Conceptual Proposal (or Philosophical Paper):* Structures an area in the form of a taxonomy or conceptual framework, hence provides a new way of looking at existing things.
- *Experience Report:* Includes the experience by the author on what and how something happened in practice.
- *Opinion Paper:* Personal opinion on a special matter is discussed in an opinion paper without relying on related work and research methodologies.

However, the two mapping studies use different ways of classifying the research focus. Basically, the research questions within each mapping study drive the actual structuring of the topic, but not the overall objective to create a systematic map of the status of the research related to software product line testing. Thus, the two maps ought to be comparable given the overarching goal of systematic mapping studies, see definition in Section 1. The different ways of structuring the topic make it a little hard to compare the conclusions in relation to the technical aspects presented in the two studies. However, a mapping between them is presented in Section 5 as part of the results.

Next, the two studies are summarized briefly in terms of approach taken to conduct each mapping study.

4.2 Swedish study

The Swedish study [Engström11] has four research questions focusing on challenges, publication forums, topics of research and type of research, and the researchers searched for publications “with a clear focus on the testing of a software product line”. The study includes only peer-reviewed articles up until 2008 (inclusive). To identify relevant papers, the researchers used a five-step search strategy:

1. Exploratory search: Six known papers were scanned for references to and from additional articles. This resulted in having 24 articles.
2. Related work: For each of the papers in step 1, introductory sections and related work sections were read and an additional 10 papers were included based on references.

3. Conference proceedings: Based on the publication forums for the papers in steps 1 and 2, the proceedings of two conferences were searched. This added another 19 papers to the set of studies.
4. Databases: The set of papers identified were validated against keyword searches in GoogleScholar and the ISI database, using general keywords as ‘‘product” and ‘‘line/ lines/family/families” and ‘‘test/testing”. This resulted in an additional 11 papers.
5. Earlier review: The earlier Spanish study was found [Lamancha09] and it was checked whether some papers were missing in the study. No new papers were identified.

The authors decided to only include peer-reviewed papers with a clear focus on some aspect of software product line testing. In total 64 papers were identified. The authors classified the papers into three main categories, which were structured as follows:

- Contribution with five subcategories: tool, method, model, metric and open items.
- Research focus with seven subcategories: organization and process, test management, testability, automation, unit testing, integration testing, as well as system and acceptance testing.
- Research type with six subcategories (as defined above): experience report, opinion paper, conceptual proposal, solution proposal, validation research, and evaluation research.

The papers were classified into the above subcategories by two researchers working independently and if there was any disagreement it was discussed and resolved. A single paper may be classified as having several different contributions and hence the sum of the contributions does not become 64. The sum reported is 74 for contributions. Furthermore, one paper has two research foci and hence the sum of research focus and research type both sums to 65.

4.3 Brazilian/USA study

The Brazilian/USA study [Neto11] uses nine specific research questions to structure the research topic. The research questions related to: test strategy, static and dynamic analysis, testing level, regression testing, non-functional requirements, commonality and variability, binding times and testability, test effort reduction, and measurement. The paper also reports where papers related to testing of software product lines are published. The study includes papers up until 2009 (inclusive). The study is primarily focused on peer-reviewed articles, but does include some book chapters, reports and theses found on the web. The non-peer-reviewed sources were primarily identified through a snowballing approach, i.e. using the reference lists of the articles identified to find more sources. The following search strategy was used:

1. A preliminary search was conducted with a set of keywords. However, it generated too many results and too few relevant results.
2. Based on the experience from step 1, the search strings were rephrased. It resulted in 18 different search strings, originating from the specified research questions.
3. The search strings were used to search a number of the major databases for software engineering publications: ScienceDirect, Scopus, IEEE Xplore, ACM Digital Library and Springer Link.
4. The search in step 3 was then complemented with a search targeting some of the major journals published in the area. This included journals published by Elsevier, IEEE, ACM and Springer.

5. A targeted search was also conducted on some of the major conferences expected to publish articles related to the topics of the systematic mapping study.
6. Finally, the reference lists of the articles found were studied to identify additional sources of information such as book chapters, technical reports and theses.

The searches in steps 4 and 5 found many of the articles already found from step 3. In total 120 papers were investigated and after having scrutinized them in two steps the final set of papers was identified. The first evaluation was done based on abstract and conclusions and the second evaluation consisted of a full paper evaluation. Finally, 45 papers were identified for inclusion in the systematic map. The authors used three inclusion criteria to decide whether a paper should be part of their study or not. The papers should include one of three aspects related to software product line testing: testing concerns (e.g. methods, techniques and testability), static and dynamic analysis, and testing effort.

The authors classified the papers in relation to the nine research questions as well as in relation to research type, resulting in the following categories:

- Research focus with subcategories: test strategy, static and dynamic analysis, testing level, regression testing, non-functional requirements, commonality and variability, binding times and testability, test effort reduction, and measurement.
- Research type with six subcategories: experience report, opinion paper, philosophical papers[2], solution proposal, validation research, and evaluation research.

As with the Swedish study, one paper may be classified into several subcategories. Each paper was categorised by two people, who then discussed and resolved any disagreements. It is worth noting that the research focus in the Brazilian/USA study has clear parallels with the research focus in the Swedish study, although covering some aspects not mapped out in the Swedish study. The contribution dimension used in the Swedish study partially relates to the Brazilian/USA study, for example, it is reasonable to compare metrics in the contribution in the Swedish study with measurement in the Brazilian/USA study. The mapping between the studies is further elaborated in Section 5. Furthermore, the Brazilian/USA study conducted a quality evaluation of the papers. The evaluation was not used to include or exclude papers, although a quality grading is provided [Neto11].

Given that the two mapping studies came up with two different sets of papers, although with some overlap, and that the classifications differ, at least in part, the only meaningful quantitative comparison between the two mapping studies is to look at the percentages of papers in different subcategories.

[2] We will henceforth use the term ”conceptual proposals” as in the Swedish study. The two studies use different names but refer to the same definition by Wieringa [Wieringa06].

5. Results

5.1 Papers

From simply listing the papers in the two mapping studies, it is possible to get a perception of the overlap between papers in the two mapping studies. The Swedish study and the Brazilian/USA study have 33 papers in common (Appendix A1), which implies that the Swedish study has 31 unique papers in comparison to the Brazilian/USA study and the Brazilian/USA study has 12 unique papers in relation to the Swedish study.

The papers for the two mapping studies are listed in Appendix A. It is noteworthy that almost 50% of the papers in the Swedish study do not appear in the Brazilian/USA study and five papers included in the Brazilian/USA study within the joint scope and time frame are not included in the Swedish study. This points to some interesting analyses, i.e. despite the same focus (testing of software product lines), there are clear differences in which papers are included. Some of these differences may be due to different years being covered, different inclusion/exclusion criteria, different search strategies or different judgments. Thus, it calls for a more detailed investigation into the differences between the studies and the papers included.

Beginning with the research questions, the two studies have different phrasing of their research questions, which also had a direct impact on their search and classification procedures. While the Swedish study was driven by a relatively high-level and open goal of identifying challenges, publication forums, topics of research and type of research on the testing of a software product line, the Brazilian/USA had nine more specific research questions, composed of the following keywords as well as their similar nouns and syntactic variations (e.g. plural form): Verification, Validation; Product Line, Product Family; Static Analysis, Dynamic Analysis; Variability, Commonality, Binding; Test Level; Test Effort, Test Measure; Non-functional Testing; Regression Testing, Test Automation, Testing Frame- work, Performance, Security, Evaluation, Validation. This list of keywords indicates that the Brazilian/USA study has a more inclusive definition of the field of software product line testing. For example, performance and security (without the testing qualifier), verification, and static analysis are not included in the Swedish study definition of software product line testing. This calls for further analysis of the consequences of these differences.

To further study the inclusion/exclusion of papers, the authors of the two systematic maps [Engström11, Neto11] were asked to identify when in the research process the papers not present in their own study (Swedish or Brazilian/USA), but included in the other study, were either not found or excluded. If a paper was excluded, the authors were asked to provide information about in which of the steps each paper was excluded. A summary of the results regarding which papers are included and excluded in the Swedish study and the Brazilian/USA study respectively can be found in Table 1.

Brazilian/USA study vs. Swedish study – 31 papers (Appendix A2)

First of all it should be noted that all papers in the Swedish study could potentially be included in the Brazilian/USA study. The Brazilian/USA study covers all years covered by the Swedish study, it includes peer-reviewed papers as focused on in the Swedish study, and has the more inclusive definition of the field. However, the specific phrasing of research questions may have constrained the scope of the Brazilian/USA study. The 31 papers included in the Swedish study and not included in the Brazilian/USA study are listed in Appendix A2. Out of these 31 papers, 21 papers were not

found in the searches and 10 papers were excluded based on the full text (Appendix A2.1). The relationships between the papers in the Swedish study and the Brazilian/USA study are summarized in Table 1. The 21 papers are further explored below.

Swedish study vs. Brazilian/USA study – five papers (Appendix A3.1)

The Brazilian/USA study includes five papers that are not included in the Swedish study, which are peer-reviewed and published in 2008 or earlier. These five papers are listed in Appendix A3.1. Four of these papers were not found in the Swedish study and one paper was excluded since it was judged to be too similar to another paper. It was judged in the Swedish study that inclusion of this paper would have biased the statistics when it came to the classification of the studies.

Summary of papers included in the two systematic maps

The outcome of the comparison between the Swedish study and the Brazilian/USA study in terms of papers included/excluded is summarized in Table 1.

Table 1: A summary of the results regarding papers in the two systematic maps

	Brazilian/USA	Swedish	Comments
Included	45	64	Differences in years and inclusion/exclusion criteria
Common	33	33	Peer-reviewed papers published 2008 or before
Not included in other study	5	31	Papers not included may be so for two reasons:
- Not found in other study	4	21	1. Papers are not found due to search strategy
- Excluded in other study	1	10	2. Papers are excluded due to different judgments
Not possible to find in other study	7	0	Papers published after 2008, or non-peer-reviewed
Potentially in common	33+1	33+10	33+10+1 = 44

It may be observed that in total the two mapping studies could potentially have 44 papers in common, but due to different judgments of papers only 33 of these papers were actually included in both studies. This means that for 25% of the papers that the studies might have had in common were excluded by one of the mapping studies.

The fact that different papers are found is also interesting. Out of the four papers found in the Brazilian/USA study and not found in the Swedish study (due to search strategy), two are a consequence of different definitions of the field. They do not mention the term product lines (65), or are about static verification (67), which is outside the Swedish definition of the field. One paper (68) is probably neglected in the title-based search, since the title is very general, while the content is specific to software product lines. Finally, one paper (69) was overlooked by mistake. It was not as straightforward to understand why 21 papers found in the Swedish study were not found in the Brazilian/USA study. Thus, these papers were studied in more detail. The following procedure was applied:

1. Search for each paper in GoogleScholar with the objective of identifying the publisher, in particular to identify whether the paper is actually available in one of the databases used for the Brazilian/USA study.
    a. Search for all words in title
    b. Search for all words in title and authors
    c. Search for exact title

2. Based on the findings classify papers into:
    a. Available in the databases used in the Brazilian /USA study – Go to database
    b. Available in any other database – Go to database
    c. Pdf available through GoogleScholar
    d. Paper only visible through citations

The above procedure resulted in the classification in Table 2.

Table 2: Classification according to GoogleScholar search of papers found in snowballing, but not in the Brazilian/US study database searches.

Class	In databases	In other databases	Pdf available	Only cited	Not found
Number of papers	6	1	12	1	1

The results in Table 2 show that all 21 papers can be found through GoogleScholar, although two papers are not available as such. Having said this, it is important to note that even searching for all words in the title generates a large number of results, which means that it would be very difficult to find the papers. It is definitely a challenge to identify search strings in GoogleScholar that lead to high precision and high recall. Further, the content of GoogleScholar changes more dynamically over time and in a less controlled manner than, for databases such as ISI Web of Science. Despite these weaknesses, GoogleScholar is an important tool and it has a very broad coverage of published research. Given that the main search strategy in the Brazilian/USA study was to search through well-established scientific databases, it is particularly interesting to look closer at the six papers actually available in the databases used. The six papers are distributed as follows across databases:

- IEEE Xplore: 3 – Papers 44, 45 and 46 in Appendix A2.2.
- SpringerLink: 2 – Papers 47 and 48 in Appendix A2.2.
- ACM Digital Library: 1 – Paper 49 in Appendix A2.2.

These six papers were most likely not found due to use of different words in the papers and the actual search strings used in the Brazilian/USA study. The papers are listed in Appendix A2.2. The six papers could be compared with the four papers not found in the Swedish study, but found in the Brazilian/USA study. It means that the number of papers not found, but possible to find in the two studies is comparable. Having said that, 16 papers are not possible to find in the databases used in the Brazilian/USA study. These 16 papers are listed in Appendixes A2.3-A2.6.

Reflections based on findings

This analysis illustrates some of the difficulties related to inclusion/exclusion criteria and that the researchers may be influenced by their research questions. The latter refers to the fact that the primary goal of mapping studies is to provide a broad review of an area (as included in the definition in Section 1), and hence a systematic mapping should not be driven by specific research questions to the same extent as a systematic literature review, which normally is more focused. However, it is very likely that search strings are formulated in relation to the research questions actually posed in a systematic mapping study even if the objective of a mapping study is to be broader.

The analyses of the two studies illustrate that it is highly likely that there will be differences in terms of papers included for different studies. Further, the analysis of the availability of papers illustrates

that researchers are highly likely to miss a number of papers because they are not available in the databases normally used when conducting database-based systematic literature reviews.

Here, we would like to hypothesize that these difficulties are because we can identify different sets of papers.

- Actual population (sometimes called the gold standard [Dieste09, Zhang11]): the set of all papers in the area based on the definition of it by the researchers conducting the secondary study. Furthermore, this is delimited by factual limitations, such as for example, publication years.
- Study population: the set is bounded by a number of decisions relating to conduct of the study, for example, databases searched.
- Sample: the actual sample in a study, which is the outcome of, for example, making judgments in relation to inclusion and exclusion criteria.

These three sets of papers are related to each other as follows. The study population is a subset of the actual population, and the sample is a subset of the study population. This is further elaborated in Section 6. Independently, we will most likely have slightly different outcomes depending on the researchers classifying the papers. This may be acceptable if the subset included in each study results in the same conclusions about the status of the area.

5.2 Research type

The researchers undertaking each of the systematic mapping studies chose to use the research type classifications proposed by Wieringa et al. [Wieringa06] and briefly introduced in Section 4.1. This simplifies the comparison between the studies. Note that the authors of the Swedish study used the term "conceptual proposal" while the Brazilian/USA study used the term "philosophical paper", as did Wieringa et al. originally. The main motivation for the change being that the Swedish study authors thought it was a better description of what Wieringa describes as papers that describe new conceptual frameworks. Here, we have chosen to follow the Swedish study and use the term "conceptual proposal".

Distributions of the classifications for the two systematic mapping studies using the above six research types are presented in Table 3. The actual classifications of the joint papers can be found in Appendix B.

Table 3: Classification of research types in the two studies.

	Distribution of all papers		Distribution of joint papers	
Research type	Swedish	Brazilian/USA	Swedish	Brazilian/USA
Evaluation research	10 (15%)	5 (11%)	8 (24%)	5 (15%)
Validation research	12 (19%)	4 (9%)	8 (24%)	3 (9%)
Solution proposal	26 (40%)	26 (58%)	10 (30%)	19 (58%)
Conceptual proposal	11 (17%)	2 (4%)	5 (15%)	1 (3%)
Experience report	2 (3%)	4 (9%)	2 (6%)	4 (12%)
Opinion paper	4 (6%)	4 (9%)	0 (0%)	1 (3%)
Total	65	45	33	33

Despite using the same classification framework, the two mapping studies paint quite different pictures of the research types used in testing of software product lines. The studies agree that solution proposal dominates the literature. They are also in agreement about the number of experience reports and opinion papers being quite small. However for all other research types there are quite big disagreements, for example, the studies disagree about whether evaluation or validation research is largest. The studies are also in quite a disagreement about the conceptual proposal papers.

The disagreements called for a more in-depth study. The mapping of papers to the research type can be found in the original publication [Engström11] for the Swedish study. For the Brazilian/USA study, the authors provided the mapping to the first author of this paper. It turns out that only 11 papers out of 33 papers are classified as the same research type in the two studies, as can be seen in Appendix B. Although each paper could be classified as one of six types, making perfect agreement close to impossible, it is still discouraging that only 33% of the papers are classified in the same way by the two research teams producing the systematic maps. Given the low level of agreement in the classification, it was judged that a calculation of inter-rater agreement as conducted by, for example, Henningsson and Wohlin [Henningsson04] was not very useful, instead it was decided to study the differences in more depth.

We looked to see if there were any systematic patterns, for example, one of the research groups classifying a paper in one type and the other group often classifying it in another type, and hence the 22 papers classified differently are studied in more detail. It results in the 22 papers being spread across 11 different pairs of classifications. If any pattern can be discerned, it is that the Swedish study has a tendency to classify papers into evaluation or validation when the Brazilian/USA study classifies papers into the solution type. This pattern is visible for 10 papers out of the 22 papers classified differently by the two groups. It potentially illustrates that these types are not necessarily disjoint, since a solution may be proposed and then either evaluated or validated. It then becomes up to the researchers to decide which of these is most prevalent in the paper, or to classify papers into several types. The observation that there seems to be some underlying pattern makes the differences in classifications somewhat less critical, although there are still opportunities for improvement. The classification differences are further discussed in Section 6.

In summary, the results indicate dissimilarities in both selection and classification, which threatens the reliability of the systematic maps and reviews. The basic idea is of course that findings from systematic maps and reviews should be similar if targeting the same topic. This is a prerequisite to build trust in systematic maps and reviews. Having said this, the findings are contradictory to the findings by MacDonell et al. [MacDonell10]. They found that systematic reviews are a robust research method. However, the contexts are different. The comparison reported here is based on two independent teams studying the same area and a broader topic than addressed by MacDonell et al. In [MacDonell10], a comparison of two teams with a joint starting point and similar experiences came to similar conclusions although approaching the systematic review in different ways. The findings here are more aligned with the results by Kitchenham et al. [Kitchenham12a], where the authors found overlap in papers, but not as high a degree of overlap as in the study by MacDonell et al.

5.3 Research focus

The two mapping studies have used different ways of classifying the research of the papers. The Swedish study has used two dimensions: contribution and research focus. For the contribution category, the study used five subcategories and for research focus, seven subcategories were used. The Brazilian/USA study used nine subcategories for research focus. First of all, it can be observed that the classifications have been used differently in the two mapping studies. Both studies use multiple classifications for the papers, i.e. a paper may contribute to more than one area and hence it is classified into more than one subcategory. However, the Swedish study is quite restrictive with its use of multiple classifications, for example, for the research focus category only one paper is classified into more than one subcategory and for contribution the total number of classifications is 74 for 64 papers. While for the Brazilian/USA study, the 45 papers are classified into 130 subcategories with respect to research question (or focus). The latter means that the Brazilian/USA study on average used close to three subcategories for each paper while the Swedish study used slightly more than one subcategory for each paper.

Given the different classification schemes and the use of the schemes in terms of multiple classifications, it becomes close to impossible to compare the two studies regarding research focus. This is not surprising since they are mapping studies, and hence if the classes that the area is mapped to are different then the results become hard to compare. Given that they are two mapping studies, no synthesis is provided. It ought to be different if comparing two systematic literature reviews, since for this type of study a synthesis is expected.

The findings about dissimilar selection and classification in the previous section, when the two studies used the same classification scheme and still did not obtain similar results, do not encourage comparisons between the findings (with different classifications and different use of the subcategories). However, it is possible to map out the papers using the classifications in the mapping studies [Engström11, Neto11]. The results are presented in Table 4, where the main observation is related to frequencies in terms of areas having received most attention from researchers. For example, it can be seen that five papers (1, 11, 15, 16 and 33) have been related to test strategies and test management in combination. It is also worth noting that several cells are empty in Table 4.

Table 4: Classification of the research focus for the joint 33 papers. The Swedish categories are shown in columns, and the Brazilian/USA in rows.

Classification of papers (33 papers)	Test organization and process	Test management	Testability	System and acceptance testing	Integration testing	Unit testing	Automation
Test strategies	12, 17	1, 11, 15, 16, 33	18	2, 3, 14, 27, 29, 31, 32	30, 32		
Static and dynamic analysis		1, 6					
Testing levels	12, 17	1, 11, 23, 33	18	9, 13, 19, 25, 27, 32	20, 30, 32		
Regression testing	12, 17	23					
Non-functional requirements				24, 29		7	
Commonality and variability		1, 4, 11, 33	18	2, 3, 9, 13, 14, 19, 25, 28, 29, 31, 32	20, 30, 32	7	5, 22
Variant binding times and testability	12	4, 11		28			22
Test effort	21	1, 11, 33	18	2, 9, 10, 14, 19, 24, 25, 26, 28, 30, 31, 32,	20, 32	7	5, 8
Testing measures	12	1		32	32		8

5.4 Validity threats

This is an analysis of two systematic maps, and hence the generalizability is of course not large. The internal validity is viewed as being high. The main reason being that a researcher working independently from the authors of the two systematic maps has performed the main analysis, and at the same time has been able to get complementary information from the authors of the systematic maps. The analysis complements the studies by MacDonell et al. [MacDonell10] and Kitchenham et al. [Kitchenham12a]. The findings are different from those of MacDonell et al., but similar to the findings by Kitchenham et al. The difference is that in [MacDonell10] the findings support the view that systematic literature reviews are reliable and here it indicates that it is a challenge to make secondary studies complete and hence reliable. Most likely it also depends on the area of study. MacDonell et al. studied a more focused area, which most likely affected the outcome. Furthermore, the researchers started from the same research questions, which also may have affected the results in comparison to the studies by Engström and Runeson [Engström11] and Neto et al. [Neto11], although the main objective is to map out an area given the definition of systematic maps.

6. Discussion

Based on the analysis of the two systematic mapping studies, some areas for discussion have been identified. First of all, it is clear that although the ambition is to summarize all relevant research in an area, different sets of papers will be obtained given a number of decisions taken. This is discussed in

Section 6.1. Some reflections regarding the findings are presented in Section 6.2. The research presented also gave rise to some areas of concern in relation to the procedures used for conducting secondary studies. This is discussed in Section 6.3. Issues related to researchers conducting secondary studies are discussed in Section 6.4. Finally, some reflections related to a reliability comparison between this study and the findings by MacDonell et al. [MacDonell10] are provided in Section 6.5.

6.1 Populations and sample

The results illustrate that it is far from obvious that the same papers will be either found or included even if secondary studies set out to analyse the same area, i.e. in this case the area of software product line testing. Before conducting the study, the first author expected more overlap between papers included in the studies, based on the two systematic maps studying the same research topic and the findings by MacDonell et al. [MacDonell10], who found a substantial overlap in the sets of papers found by two teams. Kitchenham et al. [Kitchenham12a] expressed this as 90% accuracy (in overlap of studies), which makes the authors conclude that it may be the best we can expect from two systematic literature reviews addressing the same research questions. Thus, although the objective is to find "all" relevant research papers and hence being comparable to a census, it is more likely that a sample of relevant research papers will be obtained. The differences in findings in overlap between the study by MacDonell et al. and this one are discussed further in Section 6.5.

The differences between all relevant research papers and the fact that we do not find the same papers in secondary studies are most likely due to decisions taken in the process of performing a secondary study. We propose dividing the paper selection process into three sets of papers, where the sample is a subset of the study population, which in its turn is a subset of the actual population. The three sets are as follows:

1. Actual population
   The factual criteria of the study limit the maximum number of papers that can be included. We refer to this as being the actual population, i.e. the most complete population we could obtain if we had unlimited resources to search all literature available. The factual criteria are:
   - Definition of area – the area of the secondary study is defined. Here, it refers to papers on software product line testing. However, it should be noted that even where researchers conduct studies of the same area, they might define the area slightly differently. For example, does software product line testing include papers on formal verification and dynamic analysis? It illustrates that we may not have sufficiently strong definitions of areas in software engineering, or that existing ones [Abran04] are not always used.
   - Research type – the study may focus on different types of research, for example, it may be required that a paper should include some form of empirical evaluation to be included in the study (i.e. making it potentially possible to synthesize the findings). It is worth highlighting that for systematic literature reviews, the papers must include some form of suitable data to allow for any meaningful synthesis.
   - Years – any restrictions on the years to be included in the study.
   - Language – research is published in many different languages, although in most cases researchers are focusing on publications available in English.
   - Publication type – this includes restricting the scope to, for example, only journal papers, only peer-reviewed papers or only to papers published in a specific conference/journal, or any other restriction on publication type.

These five criteria would result in a tentative maximum population, which we refer to as the actual population. It is of course close to impossible to find all papers in this population unless the secondary study is very focused, for example, on a specific series of conferences, one journal or focusing on one specific publication year.

2. Study population

   Although an actual population exists, researchers have to make decisions to instantiate the factual criteria in practical procedures. This results in a study population. The instantiation relates to three areas:

   Search strategy

   - Researchers have to decide where to look for relevant papers. This may include:
     - Databases from publishers such as Scopus or IEEE Xplore,
     - General meta-databases as Inspec or Citeseer,
     - The use of specific search engines as, for example, GoogleScholar,
     - Selection of specific journals or conferences that are relevant for the area being studied,
     - Key authors in the area of study.
   - Snowballing – it must be decided if snowballing will be used, which may add to the completeness of the set of papers identified.
   - Keywords in searches – the researcher must identify keywords to include in search strings, validate search strings, and also triggers for having a closer look at a paper in the reference lists when applying a snowballing procedure.
   - Authors – the researcher must decide whether the authors of papers identified should be contacted.

   Inclusion/exclusion criteria

   - Focus – the instantiation of the factual criteria implies specifying more details to judge what is to be included and excluded. For example, if it is required that a paper should include some form of empirical evaluation to be included in the study, criteria must be set up about what is meant by "empirical evaluation". Is an experience report counted? An experiment with "toy" artefacts? Is the word "case study" in the title sufficient, or must some additional criteria be fulfilled.
   - Level of evaluation – there is a need to decide how papers are evaluated in the selection, i.e. whether decisions are taken based on title, keywords, abstracts, partial reading (e.g. introduction and conclusions) or full text.

   Quality evaluation criteria

   - Quality – it must be decided if any criteria will be used for evaluating and judging quality, and whether the quality evaluation should result in exclusion of papers below a certain quality threshold. It is important to find quality criteria that can be applied in a similar way across different types of papers identified.

   The criteria in the above three areas will inevitably reduce the number of papers in the study in relation to the actual population (all relevant papers on the topic). As soon as, for example, a certain set of databases has been chosen, we have put certain restrictions on what we are able to find. Thus, we obtain a study population, which is the maximum number of papers that can be found under the given restrictions.

3. Sample

   The sample is the outcome from the instantiated procedures making up the study population. It is worth commenting on some of the procedures.

   Search strategy

   - Search strings – researchers must combine the keywords identified into one or more search strings for any database search conducted. Different databases also have different limitations that must be taken into account. All too often the search function in the

databases works in different ways, and other information sources, like GoogleScholar, are not consistent over time. Thus, although trying to ensure that the search strings are the same, it may be difficult due to the different ways the search function is implemented. Furthermore, different search engines do not have the same functionality, for example, whether or not it is possible to search only in the title, or they have other technical limitations, for example, in construction of the search strings or how the searches are actually implemented.

Inclusion/exclusion criteria

- Individuals' judgment on inclusion/exclusion criteria – each criterion must be judged by one or more researchers to decide whether a specific paper should be included or not. A common procedure is to have at least two people making individual evaluation and having, for example, three grades for papers: include, maybe, or exclude. In addition, the research expertise in the area is an important aspect when deciding which paper should or should not be included [Kitchenham11a].
- Combinations of individuals' judgments – given the judgment of the individuals involved in the evaluation, the judgments must be combined. This can be done in different ways.

Quality evaluation of papers

- As for the inclusion/exclusion criteria, although with a quality focus.

The actual sample will at the end be highly dependent both on the procedure applied and the judgments of the individuals conducting the secondary study.

The three paper sets and the aspects that make them different are summarized in Table 5.

Table 5: Summary of the aspects to address for the three papers sets.

Actual population	**Study population**	**Sample**
Definition of area	Search strategy: • Where to search? • Snowballing? • Keywords in searches? • Contact authors?	Search strategy: • Construction of search string based on keywords?
Research type	Inclusion/exclusion criteria: • Focus to be included? • Level of evaluation?	Inclusion/exclusion criteria: • Procedure for individual judgment of criteria? • Procedure for combining individual judgments?
Years	Quality evaluation criteria • Thresholds?	Quality evaluation criteria • Procedure for individual judgment of criteria? • Procedure for combining individual judgments?
Language		
Publication type		

Having listed the three sets of papers and then looking at the two studies (Swedish and Brazilian/USA), it is possible to see that the actual populations are not the same. The Swedish study stops at 2008 and the Brazilian/USA study includes 2009. The Swedish study only includes peer-reviewed papers while the Brazilian/USA study also includes some non-peer-reviewed papers.

However, it is possible to identify a common set of papers based on the factual criteria and hence having a joint actual population.

If looking at the delimiting criteria, it is quite clear that the researchers have taken different decisions as briefly described in Section 4 where the studies are presented. This means that the delimiting criteria have given the different studies different study populations, which explain some of the differences observed when it comes to the papers included in the separate studies.

Finally, when looking at the sampling and the instantiation of the delimiting criteria, this is dependent upon both judgment and ways of combining both judgment and keywords. This has to be factored in when comparing the final set of papers from different studies.

Taken altogether, it is not surprising that different studies Identified different sets of papers as shown by the studies of software product line testing. Researchers conducting a secondary study, whether a systematic review or a systematic map, have to make a lot of decisions and exercise a lot of judgment.

Based on this observation, it should be noted that a larger sample may not necessarily be better; it is primarily about the representativeness of the sample and hence a smaller sample may provide a better picture of the actual status of an area. Thus, a smaller sample does not imply that the conclusions should be weaker or less reliable, although a larger representative sample is normally to be preferred. The only objective here is to explore and evaluate the reliability in conducting systematic mapping studies, not to evaluate the two individual studies. The identification of relevant studies is further discussed in [Zhang11].

6.2 Findings

As discussed in the previous section, it is not surprising that researchers end up with different sets of papers when conducting a secondary study. This is acceptable, but the key concern is whether or not they came to the same conclusions about the status of an area. The results from the analysis illustrate:

1. Even though the researchers had the same classification scheme for the research type, the classifications of the joint 33 papers came out quite differently. This may be a result of the classification needing to be more concrete and maybe there is also a need to have illustrative examples of each research type. At the same time, it is a judgment and depending on the researchers' background and expertise it may result in a certain classification bias, for example, favouring one type of research type over another type. In the end, it is a matter of improving the procedures to ensure more common classifications. This is further discussed in Section 6.3.
2. When it comes to the research focus, the authors of the two mapping studies chose different ways of classifying the research. This makes it very hard to draw any conclusions from the comparison. However, it points to a need to have some way of representing research content in the same way as having a classification of research types, although it resulted in quite different outcomes in this analysis. Anyway, a classification scheme would form an important basis for a joint understanding of classifying research. This is also further discussed in Section 6.3.

Given that the findings are related to steps that are conducted both when performing a mapping study and a systematic review, it is here hypothesized that the findings are valid for both type of

studies. These findings indicate that there is still some way to go before we can expect secondary studies on the same topic to produce reliable results. This does not mean that we cannot learn from different secondary studies; it means that we have to be aware that the outcome of a secondary study is not necessarily reproducible. Thus, on the one hand secondary studies cannot yet be viewed as fully reliable, but on the other hand not using a structured approach to collect information of an area is worse, since it becomes impossible to even evaluate the reliability of a literature study. Thus, we must continue to refine and improve the support for secondary studies.

6.3 Procedures

Currently, we have guidelines for systematic reviews [Kitchenham07] and some studies have started using the research type classification proposed by Wieringa et al. [Wieringa06]. However, the outcome of this study indicates that it is insufficient. This study calls for:

- Standardized classification scheme with an agreed interpretation
  - The scheme for research type proposed in [Wieringa06] was not developed for secondary studies in general. It was originally proposed for use in requirements engineering. Maybe it has to be revised to fit other contexts too, or there may be a need to provide more details and examples together with the classification to ensure that the interpretation becomes more coherent. Compared to criteria used in the classification of empirical evaluations by Kitchenham et al. (2012b), Wieringa et al's scheme is defined on a higher abstraction level and added detail would probably reduce ambiguity.
  - The possibility of developing a classification scheme also for research focus ought to be further investigated. It may be an extension or adaptation of the ACM computing classification scheme [ACM11], which often is perceived as quite technically oriented. Such a classification scheme could potentially include general software engineering terms that go across different research topics, for example, process, method, technique, tool, and measurement. If managing to find some common terminology for classification, it would help in making studies more comparable and hence help in evaluating the reliability of secondary studies.
- Agreement on search strategy
  It may be infeasible to identify one strategy, but it would be good to have a small set of empirically proven strategies to identify relevant papers, including both strategies starting with database searches and snowballing. The strategies should preferably identify a higher number of relevant papers and still with as little noise as possible, i.e. high precision and high recall. The identification of good strategies should be done by evaluation and comparison of search strategies as discussed by, for example, Skoglund and Runeson, Dieste et al., Zhang et al., and Jalali and Wohlin [Skoglund09, Dieste 09, Zhang11, Jalali12].
- Agreement on inclusion/exclusion strategy
  After having identified a set of papers from applying the search strategy, it is important to have a good strategy of how to identify the relevant papers. This may include recommendations on the number of people involved in the inclusion/exclusion work, scales for judging relevance, rules for merging scores from different individuals, under which circumstances it makes sense to conduct a Kappa analysis of evaluator agreement and so forth. Kitchenham et al. [Kitchenham12b] studied these issues for formal experiments, using a rather detailed quality assessment scheme. They concluded that the median of three reviewers provided most reliable results. These findings

imply that inclusion/exclusion strategies should be more precisely defined, and that it is beneficial to have three reviewers taking part in the classification if possible.

- Consistency in reporting
  To enable a better understanding of primary studies, it is very important that authors clearly differentiate between papers and studies. A paper may contain results from many studies, and a study may be reported in several papers. It is preferable that authors make this as clear as possible. Furthermore, authors of secondary studies should also report any papers that they did not include because they present the same study as another paper. It is natural to include the most comprehensive paper in relation to a specific study or studies, but it is important to also report the other papers to be able to judge completeness in terms of papers found, and hence reliability. Kitchenham et al. [Kitchenham12a] also highlight this issue.

In summary, some guidelines exist, but more are most likely needed to ensure more consistent and reliable secondary studies.

6.4 Researcher issues

Classification is a key issue, but it is not only the responsibility of the researcher conducting the secondary study. Researchers writing papers must consider writing for possible use in secondary studies and synthesis. If at least using a standardized terminology, or preferably standardized classification schemes, like the ACM computing classification system [ACM11], then authors of papers can classify their own papers and hence making synthesis easier in the long run. Using structured abstracts also would help the classification [Budgen08]. However, this is a major shift, i.e. writing for actually making synthesis easier instead of primarily writing for publishing your own research results.

Education is also an important issue both to write for synthesis and for doing the secondary studies as such. In many cases, students as part of their research education conduct systematic literature studies. This may be good, but it should preferably be done together with researchers experienced in doing secondary studies as well as knowing the area of the secondary study.

6.5 Reliability comparison

The results in our study indicate lower reliability of secondary studies compared to the study by MacDonell et al. [MacDonell10], where the two teams come up with quite similar sets of papers. Our study is more aligned with the findings by Kitchenham et al. [Kitchenham12a], where the overlap in relation to other secondary studies is smaller. Thus, the findings here may not contradict the study by MacDonell et al. [MacDonell10], but rather be a consequence of the differences between the studies.

Some of the main differences between the findings in this study and the study by MacDonell et al. that may contribute to explain the differences observed are:

- Authors in both teams in the MacDonell study [MacDonell10] have contributed substantially to the field studied (as exemplified with 5 out of 11 papers co-authored by researchers in the study). Thus, the authors are experts in the area studied, and hence they are more likely to identify relevant papers than those not being expert in the area [Kitchenham11a].
- A more narrow area was studied, which is often the case with a systematic literature review, trying to address a specific research question. Secondary studies are to a large extent similar to

qualitative research where information has to be coded and observations may very well be interpreted slightly differently by different researchers or at least be given different weight to differing observations. Thus, a secondary study of a broader area (typically a systematic mapping study) will most likely require more judgment than a more focused secondary study (typically a systematic literature review addressing a specific research question), and hence we may very well find different reliability of secondary studies without any study actually being wrong.

Several of the above potential threats to the findings by MacDonell et al. [MacDonell10] are acknowledged in their study, and hence it does not mean that one study is more correct than the other. It simply means that we must be better in understanding when a secondary study is reliable and when it can be questioned. In summary, the study presented here does not support the conclusion that mapping studies are reliable. This implies that we must find ways of making them more reliable and to understand when they are reliable.

7. Conclusions

We compared two independent mapping studies in the field of software product line testing. The outcomes with respect to the three research questions stated in Section 3.1 are as follows:

RQ1: Which steps of the mapping studies are shown to be reliable or unreliable and why?
Given that the research goals are phrased quite differently, the two studies have a clear overlap in papers, but they also have papers that were either not found in the other study or not included in the other study. One factor is how the area actually was defined and is hence also a matter of judgment. The overlap in papers is considerably lower than in the reliability study by MacDonell et al. [MacDonell10] and more aligned with the overlap found by Kitchenham et al. [Kitchenham12a]. When comparing the classification of research type for the papers found in both studies (Swedish study and Brazilian/USA study), the classification is the same only for one third of the papers, which calls for clearer definitions.

RQ2: How are the general conclusions about the topic affected by differences in the two independent studies?
The conclusions are not the same, except for the very abstract level, which states that "More validation and evaluation research is needed to provide a better foundation for SPL testing" and "additional investigation, empirical and practical, should be performed" in the two abstracts, respectively. However, the differences at a more detailed level are partially different because the two studies asked different research questions although studying the same area, and hence it is difficult to compare the studies in relation to the software product line testing. Having said this, both studies used the same classification of papers when it comes to the type of research, and they come to quite different conclusions regarding the distribution of papers in terms of research type.

RQ3: Based on the analysis, what recommendations can be done to help improve reliability of secondary studies such as systematic reviews and systematic mapping studies?
The findings do not provide one answer to this question, instead the findings helped identifying further areas of research, which is presented next in terms of conjectures that have to be evaluated and studied further in future research.

Based on this study, we conclude that the reliability of secondary studies cannot and should not be taken for granted. The comparison of the two systematic maps on software product line testing shows that the decisions taken by researchers and the judgments exercised influences the outcome both in terms of which papers are found and what the researchers conclude from their secondary studies. This is not wrong, but it must be taken into account when evaluating secondary studies. The findings here point to a number of areas where more research is needed in terms of understanding how to conduct secondary studies in software engineering. Four specific research areas related to secondary studies are pointed out below in the form of four conjectures.

1. Snowballing based on researcher expertise and knowledge of an area is more efficient than trying to find optimal search strings. Snowballing gives more relevant papers and less noise and the expertise of the researchers is made better use of. Snowballing is recommended in Information Systems [Webster02]. The conjecture is based on that the Swedish study found more papers than the Brazilian/USA study. This is contradicted for two out of three secondary studies in software engineering [Skoglund09] when using a strictly formalized snowballing procedure. On the other hand, another study showed that although not exactly the same papers were found using database search and snowballing, the actual findings from the two search approaches were comparable [Jalali12]. More research is needed to understand which search strategy is best under which circumstances.
2. Secondary studies will not find the same papers (as in the case studied here and also according to Skoglund and Runeson [Skoglund09]) unless it is a study of a relatively narrow area with experts in the area conducting the study [MacDonell10]. Research is needed to understand when it can be expected that two independent secondary studies will find the same set of papers.
3. Secondary studies may come to the same general conclusions regarding an area even if the papers found are not the same. There is a need to identify under which circumstances the general findings from secondary studies come to similar findings although the set of identified papers are not the same.
4. Secondary studies are not reliable per se; they are highly dependent on the context of the secondary study, for example the area studied, researchers conducting the study, search approach and data available from the primary studies. Research is needed to understand the influence of contextual factors when conducting secondary studies.

Secondary studies have many aspects in common with both quantitative research and qualitative research. The definition of an area and searching for papers resemble quantitative research, but a secondary study also includes a lot of judgment and coding of papers, which show large similarities with qualitative research. The latter is one explanatory factor for several of the conjectures above. However, this is most likely not the only explanatory factor and hence more research is needed to fully benefit from secondary studies. At the end, secondary studies must be as reliable as possible so that other researchers can use them as a basis for their future research and practitioners can use the findings to take more informed decisions about what works and what does not work in software engineering in a given context. This may be very difficult to achieve, but it should definitely be the vision for secondary studies in software engineering.

Acknowledgment

We would like to thank John D. McGregor and Silvio Romero de Lemos Meira for the contributions to the original Brazilian/USA study. Furthermore, we would like to express our gratitude for the valuable feedback in relation to a seminar at University of New South Wales, Sydney, Australia. In particular, we would like to thank Dr. Aybüke Aurum and Sebastian Böll for valuable feedback that helped improve the paper. This work is part of the BESQ+ research project funded by the Knowledge Foundation (grant: 20100311) in Sweden. Finally, we would like to thank the editor and the reviewers for valuable feedback.

References

[Abran04] Alain Abran, James W. Moore, ”Guide to the Software Engineering Body of Knowledge”, IEEE Computer Society, 2004.

[ACM11] ACM Computing Classification System, accessible through: http://portal.acm.org/ccs.cfm?part=author&coll=portal&dl=GUIDE (accessed December 12, 2012)

[Basili99] V. R. Basili, F. Shull, and F. Lanubile, “Building Knowledge through Families of Experiments,” IEEE Transactions on Software Engineering, Vol. 25, No. 4, pp. 456–473, 1999.

[Boell11] S. Boell and D. Cezec-Kecmanovic, “Are Systematic Reviews Better, Less Biased and of Higher Quality?, in proceedings European Conference on Information Systems, paper 223, 2011.

[Budgen08] D. Budgen, B. Kitchenham, S. Charters, M. Turner, P. Brereton, and S. Linkman. “Presenting Software Engineering Results using Structured Abstracts: A Randomised Experiment”, Empirical Software Engineering, 13:435–468, 2008.

[Cruzes11] D. Cruzes and T. Dybå, “Research Synthesis in Software Engineering: A Tertiary Study”, Information and Software Technology, Vol. 53, No. 5, pp. 440-455, 2011.

[da Silva11] F. Q. da Silva, A. L. Santos, S. Soares, A. C. C. Franca, C. V. Monteiro and F. F. Maciel, “Six Years of Systematic Literature Reviews in Software Engineering: An Updated Tertiary Study”, Information and Software Technology, Vol. 53, No. 9, pp. 899- 913, 2011.

[Dieste09] O. Dieste, A. Griman, and N. Juristo, ”Developing Search Strategies for Detecting Relevant Experiments” Empirical Software Engineering, Vol. 14, No. 5, pp. 513–539, 2009.

[Dybå05] T. Dybå, B. Kitchenham, and M. Jørgensen, “Evidence-based Software Engineering for Practitioners,” IEEE Software, Vol. 22, No. 1, pp. 58–65, 2005.

[Ekeland10] A. G. Ekeland, A. Bowes and S. Flottorp, “Effectiveness of Telemedicine: A Systematic Review of Reviews”, International Journal of Medical Informatics, Vol. 79, No. 11, pp. 736-771, 2010.

[Engström11] E. Engström and P. Runeson, “Software Product Line Testing - A Systematic Mapping Study”, Information and Software Technology, Vol. 53 No. 1, pp. 2-13, 2011.

[Hayes99] W. Hayes, “Research Synthesis in Software Engineering: A Case for Meta-analysis,” in Proceeding 6th IEEE International Software Metrics Symposium, Boca Raton, Florida, USA, IEEE Computer Society, pp. 143–151, 1999.

[Henningsson04] K. Henningsson and C. Wohlin, “Assuring Fault Classification Agreement - An Empirical Evaluation", in Proceedings International Symposium on Empirical Software Engineering, pp. 95-104, Redondo Beach, California, USA, 2004.

[Jalali12] S. Jalali and C. Wohlin, “Systematic Literature Studies: Database Searches vs. Backward Snowballing", in Proceedings 6th International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 29-38, Lund, Sweden, September 2012.

[Kitchenham04] B. Kitchenham, T. Dybå, and M. Jørgensen, “Evidence-based Software Engineering,” in Proceedings 27th IEEE International Conference on Software Engineering (ICSE 2004), IEEE Computer Society, 2004.

[Kitchenham07] B. A. Kitchenham and S. Charters, “Guidelines for Performing Systematic Literature Reviews in Software Engineering”, Version 2.3, Technical Report, Software Engineering Group, Keele University and Department of Computer Science University of Durham, 2007.

[Kitchenham09] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey and S. Linkman, “Systematic Literature Reviews in Software Engineering - A Systematic Literature Review," Information and Software Technology, Vol. 51, No. 1, pp. 7-15, 2009.

[Kitchenham10] B. Kitchenham, R. Pretorius, D. Budgen, P. Brereton, M. Turner, M. Niazi and S. Linkman, “Systematic Literature Reviews in Software Engineering - A Tertiary Study”, Information and Software Technology, Vol. 52, No. 8, pp. 792-805, 2010.

[Kitchenham11a] B. Kitchenham, P. Brereton, Z. Li, D. Budgen and A. Burn, ”Repeatability of Systematic Literature Reviews”, in Proceedings International Conference on Evaluation and Assessment in Software Engineering (EASE). Durham, UK, 2011.

[Kitchenham11b] B. Kitchenham, D. Budgen and O. P. Brereton, “Using Mapping Studies as the Basis for Further Research – A Participant-observer Case Study”, Information and Software Technology, Vol. 53, No. 6, pp. 638-651, 2011.

[Kitchenham12a] B. Kitchenham, O. P. Brereton and D. Budgen, “Mapping Study Completeness and Reliability – A Case Study”, in Proceedings 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012), pp. 126-135, 2012.

[Kitchenham12b] B. Kitchenham, D. I. K. Sjøberg, T. Dybå, D. Pfahl, P. Brereton, D. Budgen, M. Höst and P. Runeson, “Three Empirical Studies on the Agreement of Reviewers about the Quality of Software Engineering Experiments”, Information & Software Technology Vol. 54, No. 7, pp. 804-819, 2012.

[Lamancha09] B. Pérez Lamancha, M. Polo Usaola and M. Piattini Velthius, “Software Product-line Testing - A Systematic Review”, in Proceedings of the 4th International Conference on Software and Data Technologies, pp. 23-30, Sofia, Bulgaria, 2009.

[MacDonell10] S. MacDonell, M. Shepperd, B. Kitchenham and E. Mendes, “How Reliable are Systematic Reviews in Empirical Software Engineering?”, IEEE Transactions on Software Engineering, Vol. 36, No. 5, pp. 676-687.

[Miller99] J. Miller, “Can Results from Software Engineering Experiments be Safely Combined?” in Proceedings IEEE 6th International Symposium on Software Metrics, Boca Raton, Florida, USA, IEEE Computer Society, 1999.

[Neto11] P. A. da Mota Silveira Neto, I. do Carmo Machado, J. D. McGregor, E. S. de Almeida and S. R. de Lemos Meira, “A Systematic Mapping Study of Software Product Lines Testing, Information and Software Technology, Vol. 53, No. 5, pp. 407-423, 2011.

[Petersen08] K. Petersen, R. Feldt, S. Mujtaba and M. Mattsson, “Systematic Mapping Studies in Software Engineering”, in Proceedings 12th International Conference on Evaluation and Assessment in Software Engineering (EASE). University of Bari, Italy, 26 - 27 June, 2008.

[Pickard98] L. Pickard, B. Kitchenham, and P. Jones, “Combining Empirical Results in Software Engineering,” Information & Software Technology, Vol. 40, No. 14, pp. 811–821, 1998.

[Skoglund09] M. Skoglund and P. Runeson, “Reference-based Search Strategies in Systematic Reviews”, in Proceedings 13th International Conference on Empirical Assessment & Evaluation in Software Engineering, Durham University, UK, 2009.

[Webster02] J. Webster and R. T. Watson, “Analyzing the Past to Prepare for the Future: Writing a Literature Review, MIS Quarterly, Vol. 26, No. 2, pp. xiii-xxiii, 2002.

[Wieringa06] R. Wieringa, N. Maiden, N. Mead and C. Rolland, “Requirements Engineering Paper Classification and Evaluation Criteria: A Proposal and a Discussion”, Requirements Engineering, Vol. 11, No. 1, pp. 102-107, 2006.

[Woodman12] J. Woodman, J. Thomas and K. Dickson, “How Explicable are Differences between Reviews that Appear to Address a Similar Research Question? A Review of Reviews of Physical Activity Interventions”, in Systematic Reviews, 1:37, 2012.

[Yin09] R. K. Yin, “Case Study Research – Design and Methods”, SAGE, 4th edition, 2009.

[Zhang11] H. Zhang, M. Ali Babar and P. Tell, ”Identifying Relevant Studies in Software Engineering”, Information and Software Technology, Vol. 53, No. 6, pp. 625-637, 2011.

Appendix A – Papers from the two mapping studies

A1. Papers in the Swedish study and the Brazilian/USA study

1. J. Al Dallal and P. Sorenson, “Testing Software Assets of Framework-based Product Families during Application Engineering Stage”, Journal of Software, Vol. 3, No. 5, pp. 11-25, 2008.
2. A. Bertolino and S. Gnesi, “PLUTO: A Test Methodology for Product-Families”, Proceedings 5th International Workshop Software Product-Family Engineering, Siena, Italy, 2004.
3. A. Bertolino and S. Gnesi, “Use Case-based Testing of Product-lines”, Proceedings European Software Engineering Conference / Foundations of Software Engineering, pp. 355-358, 2003.
4. M. B. Cohen, M. B. Dwyer and J. Shi, “Coverage and Adequacy in Software Product-line Testing”, Proceeding Workshop on Role of Software Architecture for Testing and Analysis, pp. 53-63, 2006.
5. A. Condron, “A Domain Approach to Test Automation of Product-lines”. Proceedings International Workshop on Software Product-line Testing, 2004.
6. C. Denger and R. Kolb, “Testing and Inspecting Reusable Product-line Components: First Empirical Results”, Proceedings 5th International Symposium on Empirical Software Engineering, pp. 184-193, 2006.
7. Y. Feng, X. Liu, and J. Kerridge, “A Product-line based Aspect-Oriented Generative Unit Testing Approach to Building Quality Components”, Proceedings 31st Annual international Computer Software and Applications Conference (COMPSAC), 2007.
8. U. Ganesan, Maurer, M. Ochs, B. Snoek and M. Verlage, ”Towards Testing Response Time of Instances of a Webbased Product-line”, Proceedings International Workshop on Software Product-line Testing (SPLiT 2005), pp. 23-34, Rennes, France, 2005.
9. J. Geppert, J. Li, F. Robler and D. M. Weiss, “Towards Generating Acceptance Tests for Product-lines”, Proceedings 8th International Conference on Software Reuse, Madrid, Spain, 2004.
10. J. Hartmann, M. Vieira and A. Ruder, “A UML-based Approach for Validating Product-lines”, Proceedings International Workshop on Software Product-line Testing (SPLiT), pp. 58-64, Boston, USA, 2004.
11. M. Jaring, R.L Krikhaar and J. Bosch, ”Modeling Variability and Testability Interaction in Software Product-line Engineering”, Proceedings Seventh International Conference on Composition - Based Software Systems, pp. 120-129, 2008.
12. L. Jin-hua, L. Qiong and L. Jing, “The W-Model for Testing Software Product-lines”, Proceedings International Symposium on. Computer Science and Computational Technology, 2008.
13. E. Kamsties, K. Pohl, S. Reis and A. Reuys, “Testing Variabilities in Use Case Models, Proceedings 5th International Workshop on Software Product-Family Engineering, pp. 6-18, Siena, Italy, 2003.
14. S. Kang, J. Lee, M. Kim, and W Lee, “Towards a Formal Framework for Product-line Test Development” Proceedings 7th IEEE international Conference on Computer and information Technology, pp. 921-926, 2007.
15. R. Kauppinen, J. Taina, and A. Tevanlinna, “Hook and Template Coverage Criteria for Testing Framework-based Software Product Families”, Proceedings International Workshop on Software Product-line Testing, pp. 7-12, 2004.
16. R. Kolb, “A Risk Driven Approach for Efficiently Testing Software Product-lines”, Proceedings 5th GPCE Young Researches Workshop, Erfurt, Germany, 2003.
17. R. Kolb and D. Muthig, “Challenges in Testing Software Product-lines”, Proceedings CONQUEST'03, pp. 81-95, Nuremberg, Germany, 2003.

18. R. Kolb and D. Muthig, “Making Testing Product-lines More Efficient by Improving the Testability of Product-line Architectures”, Proceedings Workshop on Role of Software Architecture for Testing and Analysis, pp. 22-27, Portland, Maine, USA, 2006.
19. J. J. Li, B. Geppert, F. Roessler and D. M. Weiss, “Reuse Execution Traces to Reduce Testing of Product-lines”, Proceedings International Workshop on Software Product-line Testing, 2007.
20. J. J. Li, D. M. Weiss and J. H. Slye, “Automatic Integration Test Generation from Unit Tests of EXVantage Product Family”, Proceedings International Workshop on Software Product-line Testing, 2007.
21. J. D. McGregor, “Structuring Test Assets in a Product-line Effort”, Proceedings Second International Workshop on Software Product-lines: Economics, Architectures, and Implications, pp. 89-92, 2001.
22. J. D McGregor, P. Sodhani and S. Madhavapeddi, “Testing Variability in a Software Product-line”, Proceedings International Workshop on Software Product-line Testing, pp. 45-50, 2004.
23. H. Muccini and A. van der Hoek, “Towards Testing Product-line Architectures”, Electronic Notes in Theoretical Computer Science 82 No. 6, 2003.
24. C. Nebut, F. Fleurey, Y. L. Traon, and J.-M. Jézéquel, “A Requirement-based Approach to Test Product Families”, International Workshop on Product Family Engineering, 2003.
25. C. Nebut, Y. Le Traon and J. M. Jézéquel, “System Testing of Product-lines: From Requirements to Test Cases. pp. 447–477, in Software Product-lines, Research Issues in Engineering and Management, Springer, 2006.
26. E. M. Olimpiew and H. Gomaa, “Model-based Testing for Applications Derived from Software Product-lines”, Proceedings 1st Workshop on Advances in Model-Based Software Testing, 2005.
27. E. M. Olimpiew and H. Gomaa, “Reusable System Tests for Applications Derived from Software Product-lines”, International Workshop on Software Product-line Testing (SPLiT 2005), pp. 8-15. 2005.
28. K. Pohl and A. Metzger, “Software Product-line Testing”. Communications of ACM, Vol. 49, No. 12, pp. 78-81, 2006.
29. S. Reis, A. Metzger and K. Pohl, “A Reuse Technique for Performance Testing of Software Product-lines”, Proceedings International Workshop on Software Product-line Testing, pp. 5-10, 2006.
30. S. Reis, A. Metzger and K. Pohl, “Integration Testing in Software Product-line Engineering: A Model-Based Technique”, Proceedings 10th International Conference on Fundamental Approaches to Software Engineering, Braga Portugal, 2007.
31. A. Reuys, E. Kamsties, K. Pohl, and S. Reis, “Model-based System Testing of Software Product Families”, Proceedings 17th Conference on Advanced Information Systems Engineering, pp. 519-534, Porto, Portugal, 2005.
32. A. Reuys, S. Reis, E. Kamsties and K. Pohl, “The ScenTED Method for Testing Software Product-lines”, Proceedings Software Product-lines, pp. 479-520, 2006.
33. H. Zeng, W. Zhang and D. Rine, “Analysis of Testing Effort by Using Core Assets in Software Product-line Testing”, Proceedings International Workshop on Software Product-line Testing, SPLiT, 2004.

A2. Papers only in the Swedish study

A number of papers were excluded in the Brazilian/USA, which was included in the Swedish study. These are listed in Appendix A2.1. Only six papers included in the Swedish study, where actually available in the databases used in the Brazilian/USA study, see Appendix A2.2. In Appendixes A2.3-A2.6, papers included in the Swedish study but not really possible to find in the Brazilian/USA study, due to the focus on a defined set of databases, are listed. The reasons for not finding them include papers being in other databases than those used in the Brazilian/USA study or not being available in a specific database, but possible to find on the web.

A2.1. Excluded based on full papers

34. S. Bashardoust-Tajali and J-P. Corriveau, “On Extracting Tests from a Testable Model in the Context of Domain Engineering”, Proceedings 13th IEEE International Conference on Engineering of Complex Computer Systems, pp.98-107, 2008.
35. Ganesan, J. Knodel, R. Kolb, U. Haury and G. Meier, ”Comparing Costs and Benefits of Different Test Strategies for a Software Product-line: a Study from Testo AG, Proceedings Software Product-line Conference (SPLC), 2007.
36. T. Kishi, and N. Noda, “Design Testing for Product-line Development based on Test Scenarios”. Proceedings Software Product-line Testing Workshop (SPLiT), Boston, MA, USA, 2004.
37. P. Knauber and W. Hetrick, “Product-line Testing and Product-line Development - Variations on a Common Theme”, Proceedings International Workshop on Software Product-line Testing (SPLiT), 2005.
38. P. Knauber and J. Schneider, “Tracing Variability from Implementation to Test Using Aspect-Oriented Programming”, Proceedings International Workshop on Software Product-line Testing SPLiT, 2004.
39. J. D. McGregor and K. Im, “The Implications of Variation for Testing in a Software Product-line”, Proceedings International Workshop on Software Product-line Testing, (SPLiT), 2007.
40. S. Mishra, “Specification based Software Product-line Testing: A Case Study”, Proceedings Concurrency: Specification and Programming Workshop, pp. 243-254, 2006.
41. C. Nebut, S. Pickin, Y. Le Traon, and J. M. Jezequel, “Automated Requirements-based Generation of Test Cases for Product Families”, Proceedings 18th IEEE International Conference on Automated Software Engineering, 2003.
42. C. Nebut, S. Pickin, Y. Le Traon, and J. M. Jezequel, “Reusable Test Requirements for UML-Model Product-lines”, Proceedings International Workshop on Requirements Engineering for Product-lines, 2002.
43. J. J. Williams and I. Cummins, “Test Case Management of Controls Product-line Points of Variability”, Proceedings International Workshop on Software Product-line Testing (SPLiT), 2004.

A2.2. In databases explored in the Brazilian/USA study

44. T. Kahsai, M. Roggenbach and B.-H. Schlinglof, “Specification-based Testing for Software Product-lines”, Proceedings Sixth IEEE International Conference on Software Engineering and Formal Methods, Cape Town, South Africa, 2008. (IEEE)
45. K. D. Scheidemann, “Optimizing the Selection of Representative Configurations in Verification of Evolving Product-lines of Distributed Embedded Systems”, Proceedings 10th International Software Product-line Conference (SPLC’06), pp. 75-84, 2006. (IEEE)

46. E. Uzuncaova, D. Garcia, S. Khurshid, and D. Batory, “Testing Software Product-lines using Incremental Test Generation”, Proceedings International Symposium on Software Reliability Engineering, 2008. (IEEE)
47. J. Weingärtner, “Product Family Engineering and Testing in the Medical Domain — Validation Aspects”, Proceedings 4th International Workshop on Software Product-Family Engineering, pp. 56-77, Bilbao, Spain, 2001. (Springer)
48. A. Bertolino, A. Fantechi, S. Gnesi and G. Lami, “Product-line Use Cases: Scenario-Based Specification and Testing of Requirements”, in Software Product-lines Research Issues in Engineering and Management, (Eds.) T. Käkölä and J. C. Duenas, Springer, 2006. (Springer)
49. A. Tevanlinna, J. Taina and R. Kauppinen, “Product Family Testing: a Survey”, ACM SIGSOFT Software Engineering Notes, Vol. 29, No. 2, pp. 12-17, 2004. (ACM)

A2.3. In other databases

50. R. Kolb and D. Muthig, “Techniques and Strategies for Testing Component-Based Software and Product-lines”, in Development of Component-Based Information Systems. Advances in Management Information Systems Volume 2 / 2006, pp. 123-139, 2006.

A2.4. Available in pdf through GoogleScholar

51. J. C. Dueñas, J. Mellado, R. Cerón, J. L. Arciniegas, J. L. Ruiz and R. Capilla, “Model Driven Testing in Product Family Context”, Proceedings First European Workshop on Model Driven Architecture with Emphasis on Industrial Application, 2004.
52. U. Dowie, N. Gellner, S. Hanssen, A. Helferich, G. Herzwurm and S. Schockert, “Quality Assurance of Integrated Business Software: An Approach to Testing Software Product-lines”, Proceedings 13th European Conference on Information Systems, 2005.
53. Y. Ghanam, S. Park, and F. A. Maurer, “A Test-Driven Approach to Establishing & Managing Agile Product-lines”, Proceedings International Workshop on Software Product-line Testing, 2008.
54. J. D. McGregor, “Toward a Fault Model for Software Product-lines”, Proceedings Fifth International Workshop on Software Product-line Testing, (SPLiT) 2008.
55. M. Olimpiew and H. Gomaa, “Model-based Test Design for Software Product-lines”, Proceedings International Workshop on Software Product-line Testing (SPLiT), 2008.
56. S. Oster, A. Schürr and I. Weisemöller, “Towards Software Product-line Testing using Story Driven Modeling”, Proceedings 6th International Fujaba Days, pp. 48-55, 2008.
57. A. Reuys, S.Reis, E.Kamsties and K. Pohl, “Derivation of Domain Test Scenarios from Activity Diagrams”, Proceedings International Workshop on Product-line Engineering The Early Steps: Planning, Modeling, and Managing, 2003.
58. C. Shaulis, “Salion's Quality Confident Approach to Testing Software Product-lines”, Proceedings International Conference on Product-line Testing, Boston, MA, USA (SPLiT 04), 2004.
59. Z. Stephenson, Y. Zhan, J. Clark, and J. McDermid, “Test Data Generation for Product-lines - A Mutation Testing Approach”, International Workshop on Software Product-line Testing (SPLiT), 2004.
60. A. Tevanlinna, “Product Family Testing with RITA”, Proceedings Eleventh Nordic Workshop on Programming and Software Development Tools and Techniques (NWPER), pp. 251-265, 2004.
61. T. Trew, “What Design Policies must Testers Demand from Product-line Architects?”, Proceedings International Workshop on Software Product-line Testing, 2004.

62. S. Weißleder, D Sokenou and BH Schlingloff, “Reusing State Machines for Automatic Test Generation in Product-lines”, Proceedings 1st Workshop on Model-based Testing in Practice, 2008.

A2.5. Appear as cited in GoogleScholar

63. M. Olimpiew, and Gomaa, “Customizable Requirements based Test Models for Software Product-lines”, Proceedings International Workshop on Software Product-line Testing, Baltimore, MD, USA, 2006.

A2.6. Not found in GoogleScholar

64. T. Gustafsson, “An Approach for Selecting Software Product-line Instances for Testing” International Workshop on Software Product-line Testing, 2007.

A3. Papers only in the Brazilian/USA study

It should be noted that two of the papers found in the Brazilian/USA study could not be found in the Swedish study due to searching in different time frames, i.e. the Swedish study was conducted up until 2008 and the Brazilian/USA study also included 2009 (see Appendix A3.2). Furthermore, the Swedish study decided to exclude non-peer reviewed literature, which means that five papers identified in the Brazilian/USA study were excluded by definition in the Swedish study (see Appendix A3.3). This leaves five papers found in the Brazilian/USA study that were not found in the Swedish study (see Appendix A3.1).

A3.1. Unique peer-reviewed papers from 2008 and before

65. M. J. Harrold, “Architecture-based Regression Testing of Evolving Systems”, Proceedings International Workshop on Role of Architecture in Testing and Analysis, pp. 73–77, Marsala, Sicily, Italy, 1998.
66. R. Kauppinen and J. Taina, “RITA Environment for Testing Framework-based Software Product-lines”, Proceedings 8th Symposium on Programming Languages and Software Tools, pp. 58–69, Kuopio, Finland, 2003.
67. T. Kishi and N. Noda, “Formal Verification and Software Product-lines”, Communications of the ACM, Vol. 49, No. 12, pp. 73-77, 2006.
68. D. Needham and S. Jones, “A Software Fault Tree Metric”, Proceedings International Conference on Software Maintenance, pp. 401-410, Philadelphia, Pennsylvania, USA, 2006.
69. A. Wübbeke, “Towards an Efficient Reuse of Test Cases for Software Product-lines, Proceedings Software Product-line Conference, pp. 361-368, 2008.

A3.2 Papers from 2009

70. E. M. Olimpiew and H. Gomaa, “Reusable Model-based Testing”, Proceedings 11th International Conference on Software Reuse, pp. 76-85, 2009.
71. B. Prez Lamancha and M. Polo Usaola, “Towards an Automated Testing Framework to Manage Variability using the UML Testing Profile”, Proceedings Workshop on Automation of Software Test, pp. 10-17, Vancouver, Canada, 2009.

A3.3 Non-peer reviewed papers

72. J. D. McGregor, “Building Reusable Test Assets for a Product-line”, Tutorial summary paper presented in Proceedings 7th International Conference on Software Reuse, pp. 345-346, Austin, Texas, USA, 2002.

73. K. Pohl and E. Sikora, “Documenting Variability in Test Artefacts” Chapter 8 in Software Product-line Engineering written by K. Pohl, G. Böckle and F.van der Linden, pp. 149-158, 2005.
74. J. D. McGregor, “Testing a Software Product-line”, Technical Report CMU/SEI-2001-TR-022, 2001.
75. R. Kauppinen, “Testing Framework-based Software Product-lines”, Master’s thesis, Department of Computer Science, University of Helsinki, Finland, 2003.
76. O. O. Edwin, “Testing in Software Product-lines”, Master’s thesis, Department of Software Engineering and Computer Science, Blekinge Institute of Technology, Sweden, 2007.

Appendix B – Classification of papers in research type

Comparison of classification of the 33 common research articles listed in Appendix A1. Each study's classification is shown with Brazilian respectively Swedish.

Number in A1	Validation Research	Evaluation Research	Solution Proposal	Philosophical Papers	Opinion Papers	Experience Papers
1	Swedish	Brazilian	0	0	0	0
2	Brazilian, Swedish	0	0	0	0	0
3	Brazilian	0	Swedish	0	0	0
4	0	0	Brazilian, Swedish	0	0	0
5	0	0	Brazilian, Swedish	0	0	0
6	Brazilian, Swedish	0	0	0	0	0
7	Swedish	Brazilian	0	0	0	0
8	0	Swedish	Brazilian	0	0	0
9	Swedish	0	Brazilian	0	0	0
10	0	0	Brazilian, Swedish	0	0	0
11	0	Swedish	Brazilian	0	0	0
12	0	0	Brazilian, Swedish	0	0	0
13	0	0	Brazilian, Swedish	0	0	0
14	0	0	Brazilian, Swedish	0	0	0
15	0	0	Brazilian	Swedish	0	0
16	0	0	Swedish	0	Brazilian	0
17	0	0	0	0	Swedish	Brazilian
18	0	0	0	Swedish	Brazilian	0
19	0	Swedish	Brazilian	0	0	0
20	0	Swedish	Brazilian	0	0	0
21	0	Brazilian	0	Swedish	0	0
22	0	Swedish	Brazilian	0	0	0
23	0	0	Brazilian	0	Swedish	0
24	Swedish	0	Brazilian	0	0	0
25	Swedish	0	Brazilian	0	0	0
26	0	0	Brazilian, Swedish	0	0	0
27	0	0	0	Swedish	Brazilian	0
28	0	0	0	Swedish	Brazilian	0
29	0	Brazilian,	0	0	0	0

		Swedish				
30	Swedish	0	Brazilian	0	0	0
31	0	Brazilian, Swedish	0	0	0	0
32	0	Swedish	Brazilian	0	0	0
33	0	0	Swedish	Brazilian	0	0