Repeatability of systematic literature reviews

[1]Barbara Kitchenham, [1]Pearl Brereton, [2]Zhi Li, [3]David Budgen and [3]Andrew Burn

[1]School of Computing and Mathematics, Keele University, Staffordshire ST5 5BG
{O.P.Brereton,B.A.Kitchenham}@cs.keele.ac.uk
[2]College of Computer Science and Information Technology, Guangxi Normal University,
No.15 Yu Cai Road, Guilin, Guangxi 541004, P.R. China.
zhili@mailbox.gxnu.edu.cn
[3]School of Engineering and Computing Sciences, Durham University,
South Road, Durham City, DH1 3LE, UK
{David.Budgen,A.J.Burn}@durham.ac.uk

Abstract

Background: One of the anticipated benefits of systematic literature reviews (SLRs) is that they can be conducted in an auditable way to produce repeatable results. **Aim:** This study aims to identify under what conditions SLRs are likely to be stable, with respect to the primary studies selected, when used in software engineering. The conditions we investigate in this report are when novice researchers undertake searches with a common goal. **Method:** We undertook a participant-observer multi-case study to investigate the repeatability of systematic literature reviews. The "cases" in this study were the early stages, involving identification of relevant literature, of two SLRs of unit testing methods. The SLRs were performed independently by two novice researchers. The SLRs were restricted to the ACM and IEEE digital libraries for the years 1986-2005 so their results could be compared with a published expert literature review of unit testing papers. **Results:** The two SLRs selected very different papers with only six papers out of 32 in common, and both differed substantially from a published secondary study of unit testing papers finding only three of 21 papers. Of the 29 additional papers found by the novice researchers, only 10 were considered relevant. The 10 additional relevant papers would have had an impact on the results of the published study by adding three new categories to the framework and adding papers to three, otherwise empty, cells. **Conclusions:** In the case of novice researchers, having broadly the same research question will not necessarily guarantee repeatability with respect to primary studies. Systematic reviews must be careful to report their search process fully or they will not be repeatable. Missing papers can have a significant impact on the stability of the results of a secondary study.

Keywords: Systematic Literature Review, Repeatability, Case Study

1. Introduction

The EPIC (Evidence-based Practices Informing Computing) project has been investigating properties of Systematic Literature Reviews (SLR) in the Software Engineering (SE) domain. This paper reports the results for one of the research questions addressed by a case study to investigate the stability of the SLR process. The specific research question addressed is:

> RQ1: To what extent does the SLR methodology provide repeatable results?

The basic methodology used is a participant-observer multi-case study using Yin's approach to case study design (Yin 2003). The "case" is the part of the SLR conduct stage that involves identifying and selecting the relevant primary studies. In this study, the results from two SLRs carried out independently by two research assistants are compared to one another. They are also compared with an expert literature review addressing the topic of empirical studies of unit testing undertaken by Juristo and colleagues (Juristo et al., 2004 and Juristo et al., 2006), henceforth referred to as the JMV study.

Following Yin's terminology, the case study protocol defined the following propositions related to the research question:

> RQ1-P1: Where an SLR is conducted within well-defined parameters describing the topic, information sources and chronological bounds,

the researchers should find the same set of sources.

RQ1-P2: Where an SLR is repeated using the same parameters and guidelines for analysis, then the analyses will produce the same conclusions.

This report addresses both of these research propositions. The method used to address the propositions was defined in the case study protocol (Budgen 2009) and is summarised in Section 2 which also includes an overview of the three literature reviews used in the study. The results from the three literature reviews are reported in Section 3, and the extent to which they confirm or contradict the research propositions is discussed in Section 4. The conclusions are reported in Section 5.

2. Methodology

We carried out a *participant-observer* multi-case case study. The cases are the primary study identification and selection steps of a published expert literature review (the baseline case) plus two independent SLRs undertaken by research assistants.

The aim of the case study was to determine whether the process of identifying relevant studies in an SLR for a given topic, using the same guidelines, over the same time period, produces consistent results.

2.1 Choice of Topic

The topic chosen was methods for unit testing and the baseline was a previous study by Juristo *et al* (Juristo 2006). Software testing is an important software engineering task and it has a strong impact on the quality of software related products (Jørgensen 2008). According to the JMV study, there are a sufficient number of primary studies that compare unit testing methods to make an evidence-based software engineering study worthwhile. We emphasis, however that Juristo et al. did not claim that their secondary study was an SLR and we consider it to be an example of an expert literature review not an SLR.

2.2 Case Study Roles

This study was conducted entirely within the EPIC team, and hence some of the team members performed specific roles as case study researchers as well as roles in the systematic literature reviews. In the SLRs, the roles were:

Supervisors: Pearl Brereton and David Budgen

Systematic Literature Reviewers (carrying out the independent SLRs): Zhi Li (Keele) and Feng Bian (Durham)

SLR protocol reviewers: Pearl Brereton, David Budgen, Barbara Kitchenham, Stephen Linkman, Mahmood Niazi and Mark Turner.

In the case study, David Budgen was the case study leader, and he and Pearl Brereton also acted as observers, maintaining records of their supervisory activities. The other members of the case study team were: Barbara Kitchenham, Stephen Linkman, Mahmood Niazi, Michael Goldstein and Mark Turner and Andrew Burn who took over from Feng Bian when he left the EPIC project. (Note, throughout this report, individual researchers performing specific roles are named, because anonymity is impossible when accurately reporting an participant- observer study.)

2.3 Research Assistant Training

The research assistants (RAs), Zhi Li and Feng Bian, were given copies of the guidelines for performing systematic literature reviews. They also attended a post-graduate workshop on systematic literature reviews run by Kitchenham at Durham University. Note, at that time Burn did not work for the EPIC project but was a postgraduate student at Durham University and attended the workshop.

2.4 Case Study Design

The RAs carried out a direct replication of the JMV study (Juristo 2006, Juristo, Moreno and Vegas 2004) in terms of the set of digital libraries used and the time period covered. The major differences were that the RAs were instructed to use the methodology described in the SLR guidelines (Kitchenham and Charters, 2007) and, also, to adopt an automated method of searching the ACM and IEEE digital libraries.

Each of the RAs conducted a completely separate study, so that together with the original, we would have a total of three studies (two SLRs and one expert literature review). To ensure independence, neither of the RAs read the papers reporting the JMV study. Each wrote his own protocol starting by identifying the formal research question for the systematic literature review. The protocol was reviewed by the following reviewers:

- Feng Bian (Durham) reviewed by David Budgen, Barbara Kitchenham and Stephen Linkman
- Zhi Li (Keele) reviewed by Pearl Brereton, Mahmood Niazi and Mark Turner

The reviewers for each protocol were 'blinded' to the other protocol. The selection of papers and data extraction tasks were checked by the supervisors. This role was undertaken by David Budgen (Feng Bian) and Pearl Brereton (Zhi Li). The RAs were also 'blinded' as to details of the case study protocol.

These case studies addressed protocol construction and the search and selection of primary studies only. No quality assessment or data extraction was performed. The papers found by Feng Bian and Zhi Li were compared in order to address RQ1-P1. They were also compared with the papers found in JMV study since we has originally assumed that the JMV study would act as a gold standard against which to assess the search and selection processes performed in the two SLRs.

The additional papers found by Zhi Li and Feng Bian were reviewed by Brereton and Kitchenham and assessed with a refined set of inclusion/exclusion criteria. This process led to the exclusion of some of the original set of papers but identified a set of papers that we believed to be relevant to the research questions raised in the JMV study, but which were not found by that study. The primary studies reported in these papers were assessed for quality and data related to the type and outcome of each study were extracted. We also extracted quality data from the studies reported in the papers found in the JMV study. The collected data were compared with the framework for unit testing and results reported in the JMV study. This information was used to address RQ1-P2.

Table 1: Parameters of the baseline expert literature review

Factor	Stated or Derived	Values used	Comments
Research question	derived	What do we know empirically about software unit testing?	
Sources	stated	IEEE and ACM digital libraries	Referred to as an "extensive search" in Juristo et al. (2006), page 72.
Form of search	derived	Automatic	Assumed from the context of the reference to the Digital Libraries (above).
Period of search	derived	1987-2005	Juristo et al. (2004) does refer to a 1978 paper, but then discards it. The search period is not stated anywhere but the earliest paper used is published in 1987.
Inclusion	stated	Unit testing experiments (laboratory study, formal statistical analysis, laboratory replication, field study)	Discussed in both papers.
Exclusion	stated	Theoretical studies simulations	Discussed in Juristo et al. (2006) page 72
Forms of primary study	stated	Test-set generation Test-set evaluation Test-set selection	Discussed in Juristo et al. (2006) page 72
No. of primary studies used for analysis	derived	20 (2004) 24 (2006)	Not stated explicitly in either paper, so derived by counting the entries in the tables.
Form of analysis	derived	Classification from the SWEBOK chapter on testing (Chapter 5)	The use of the SWEBOK chapter is discussed more in Juristo et al. (2004). Both papers use this to classify forms of unit testing and then discuss the contributions of the (small) set of papers in each class.

Table 2: Research questions used for the independent SLRs

SLR researcher	Research question
Feng Bian	What empirical evidence has been found in the studies that assess the effectiveness and efficiency of different unit testing strategies?
Zhi Li	What empirical studies have been carried out to compare unit testing methods in software development?

2.5 The Baseline Case

The baseline expert literature review on unit testing was initially published in *Empirical Software Engineering* (Juristo et al., 2004), with a later version published in *IEEE Software* (Juristo et al., 2006). Neither paper provides much detail about the search process, though the 2006 paper is generally informative about its scope. Table 1 indicates the main parameters of the study.

2.6 The Two Replication Cases

In this section, we outline the relevant steps of the two independent SLRs undertaken by the two RAs. The RAs were both given the task of finding empirical results related to unit testing available from the ACM and IEEE digital libraries. However, the RAs were asked to formulate their own research questions and search process based on the overall goal. In each case, the research questions were formulated with the aim of aggregating the results of empirical studies that compare unit testing methods or strategies, see in Table 2.

In Feng Bian's SLR, the research question was explicitly aimed at assessing the effectiveness and efficiency of different unit testing methods. In Zhi Li's SLR, the research question explicitly targeted the comparison of different unit testing methods, although the comparison criteria were not specified.

The search process was based on automated searching of the IEEE Xplore Digital Library and the ACM Digital Library covering the time period from 1 January 1987 to 31 December 2005. This period was used so that a direct comparison could be made with the JMV study.

In Zhi Li's search, the following logical expression was used for the search terms:

((unit AND testing) OR (component AND testing)) AND (compare OR comparison OR comparing) AND (empirical OR laboratory OR experiment OR field survey) AND (software) AND ((1987 <= year) AND (year <= 2005))

In Feng Bian's search, the search context was the title and abstract of each paper, and the search period was between years 1987 and 2005. The general search strategy was searching the title and abstract by using "unit test", "unit testing", "component test" and "component testing" as the word phrase first, then adding an additional word "compare/ comparison/ effective/ effectiveness/ empirical" in turn in all fields. His search terms can be summarized by the following logical structure of search terms (for some reason, "efficiency" did not appear in his search terms):

((unit test) OR (unit testing) OR (component test) OR (component testing)) AND (effective OR effectiveness) AND (compare OR comparison) AND (empirical)

In both cases, a paper was included if:

- it described experiments (i.e., a laboratory study, formal statistical analysis, laboratory replication, or field study) that compare unit testing methods.
- it is published between January 1987 and December 2005.

A paper was excluded if:

- it only addressed theoretical concepts.
- it only described simulations.

Papers published in non-peer reviewed sources such as ACM SIGSOFT Notes were excluded unless the paper was based on a conference or workshop paper. The RAs each produced a list of candidate primary study papers based on their own inclusion/exclusion criteria.

After this initial selection process, Brereton checked the candidate papers a second time amending the inclusion/exclusion criteria as follows:

- Include case studies as long as there was a baseline for comparison. This relaxation was required because the JMV study included two case studies.
- Include regression testing papers. This was necessary because the JMV study took the decision to include regression testing papers although it could be argued that regression testing is a form of system testing
- Exclude System or Integration Testing papers (unless regression testing studies).
- Exclude testing papers for specialized types of software (such as spreadsheets).

In the case of any uncertainty, the paper was also reviewed by Kitchenham and a consensus on inclusion/exclusion was reached.

3. Results

3.1 Baseline review

The primary studies cited in Juristo et al. 2006 were essentially a super-set of those cited in Juristo et al. 2004: in a few cases, conference versions of primary studies were replaced by references to later journal versions; and 4 papers published later were added, giving a total of 24 papers in the 2006 publication. Among the 24 papers, three were not available in the ACM or IEEE digital libraries, i.e. Roper et al. (1997), Wong and Mathur (1995) and

Frankl et al. (1997). These were journal versions of conference papers found in their search of the ACM and IEEE digital libraries. Thus for purposes of comparison with searches based on IEEE and ACM digital libraries the number of papers found in the JMV set is considered to be 21.

3.2 Replication Studies Search and Selection

The numbers of papers identified and subsequently selected by the two independent reviews are summarized in Table 3. Of the selected papers, six were common to both reviews. The overlaps between the sets of papers found by the RAs are shown in Figure 1. The results shown in Figure 1 are very surprising, not only because there was so little overlap between the RAs, but because there appeared to be a large number of papers identified by the RAs but not by the JMV study (i.e. 29 papers).

The papers selected by the RAs were reviewed independently by Brereton and Kitchenham who excluded 19 of the papers. This left an additional ten papers that were apparently missed by the JMV study. This was not so startling as the initial results but still implied that the JMV study missed a relatively large number of papers.

Table 3: Number of papers found by the independent reviews

	Zhi Li			Feng Bian		
	Initial search	Screening on title and abstract	Selected	Initial search[1]	Screening on title and abstract	Selected
IEEEXplore	52	19	11	139	13	7
ACM	107	61	11	83	35	9
TOTAL	159	80	**22**	222	48	**16**

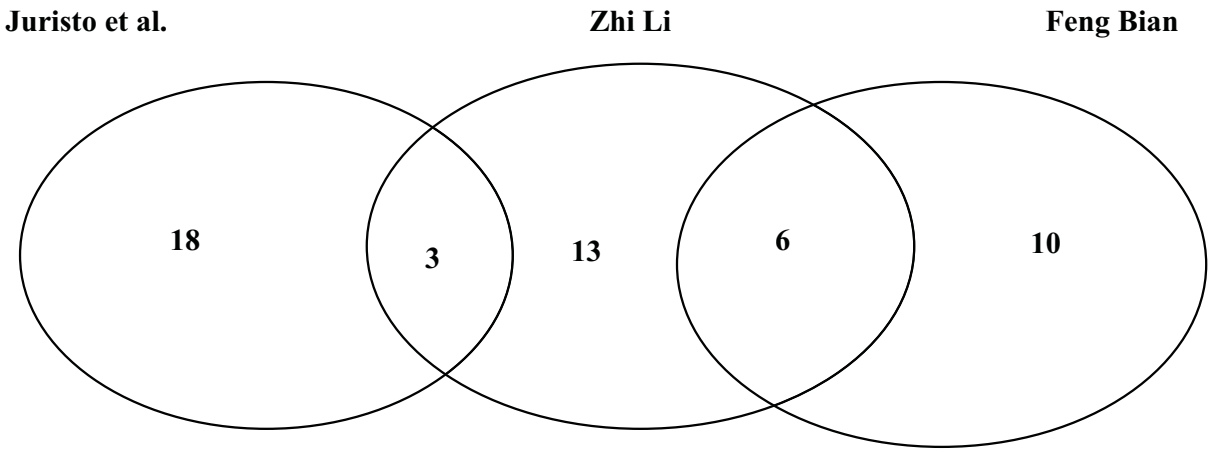

Figure 1: Overlap among papers found by RAs and the JMV study

There are several possible explanations for the lack of overlap between the different searches. For example:

- The search strings used by Zhi Li and Feng Bian did not include the term "regression testing". They had been tasked specifically to search for unit testing papers not regression testing papers. In fact, they found five regression testing papers. Strangely enough the only overlap they had with the JMV papers was based on three regression testing papers.
- As an expert review, the JMV study might have rejected poor quality papers without explicitly stating the fact. In order to investigate this possibility we undertook a quality evaluation of the papers found by our search and the papers found by the JMV study. This is examined in the next section.
- As an expert review, the JMV study might have concentrated on the most prominent/important papers. This is discussed in a later section.

[1] Including some screening of obviously irrelevant papers

3.3 Quality Evaluation

We undertook a quality evaluation of the papers found by the JMV study and the ten papers selected from the candidate papers identified by Zhi Li and Feng Bian. This process was complicated by two issues:

1. Many of the papers reported multiple studies and quality evaluation needs to take place at the study level.
2. The studies were of two very different types: human-centric and technology-centric. This meant that we found it necessary to use two different quality evaluation criteria.

Since the additional papers found by our case studies included only technology-centric papers and only three of the papers found by the JMV study were human-centric, we report only the results for the technology-centric papers. With respect to the number of studies included in the papers, the authors of the papers generally noted that they had undertaken multiple studies but we only counted studies, where there was a change to the basic experimental procedure, as a separate study. We did not count experiments that reported different aspects of the same experimental procedure as separate studies. Also we did not include "experiments" where there was no comparison (e.g. experiments that simply reported the time a process took without reference to any baseline). Thus our count of relevant experiments was sometimes fewer than the number of experiments reported by the authors. The split into individual studies was part of the quality assessment process, so it was done by two researchers and any disagreement was discussed until a consensus was reached. In all we found six papers with multiple studies. Three of the papers included in the JMV study reported eight unique studies between them. Three of the additional 10 papers reported nine unique studies.

Table 4: Quality checklist for Technology-centric papers

Question No	Question	Available answers
1	Are the study measures valid?	None=0 / Some (0.33) / Most (0.67) / All (1) Note: should score down if not actually detecting faults or changes to code.
2	Was there any replication, i.e. multiple test objects, multiple test sets?	Yes - both (1) / Yes – either (0.5)/No (0)
3	If test cases were required by the Test Treatment, how were the test cases generated?	Not applicable / By the experimenters (Yes=0) / By an independent third party (Yes=0.5) /Automatically (Yes=0.5)/ By industry practitioners when the test object was created (Yes=1)
4	How were Test Objects generated?	Small programs (Yes=0) / Derived from industrial programs but simplified (Yes=0.5) /Real industrial programs. (Yes=1)
5	How were the faults/modifications found?	Not applicable /Naturally occurring Yes=1, go to question 6 If No go to question 5a
5a	For seeded faults/modifications, how were the faults identified?	Faults introduced by the experimenters (Yes=0), / Independent third party (Yes=0.25) / Generated automatically (Yes=0.5) – inc. mutants Go to 5b.
5b	For seeded faults/modifications, were the type and number of faults introduced justified?	Type and Number (of seeds): Yes (0.5) / Type or Number (of seeds) : (Yes=0.25) / No=0 / Mutants: (Yes=0.25, No=0)
6	Did the statistical analysis match the study design?	No=(0) / Somewhat (0.33) / Mostly (0.67) / Completely (1)
7	Was any sensitivity analysis done to assess whether results were due to a specific test object or a specific type of fault? Note: look out for a further experiment which is just to check original assumptions.	Not applicable / Yes=1 / Somewhat=0.5 / No=0
8	Were limitations of the study reported either during the explanation of the study design or during the discussion of the study results?	No=0 / Somewhat=0.5 / Extensively=1

We used the quality evaluation instrument shown in Table 4. The development of the checklist was reported in Kitchenham et al. (2009). However, we made some adjustments as a result of using the form. In particular, we scored numerically and assessors were allowed to interpolate between numerical values. The checklist form also included a comment column that was used to add an explanation for the score. Papers were assigned at random to two of the four assessors (Brereton, Kitchenham, Budgen and Burn). Disagreements were discussed among pairs until a consensus was reached.The Average score for each study was calculated as follows:

Average= (TS)/(NQ-NAQ)

Where

TS = Total Score

NQ=Number of questions i.e. 8

NAQ=Number of "not applicable" questions.

We found that "not applicable" was necessary when a specific decision about the study design ruled other questions irrelevant. For example if a study just counted the number of test cases generated, then question 5 regarding the origin of faults/modification was irrelevant. We removed the number of non-applicable questions from the denominator of the average score because otherwise we would be penalizing the study design more than once. However, we note that a counter argument is that having made a bad decision that rendered other aspects of a good design "not applicable", the paper *should* be further penalized.

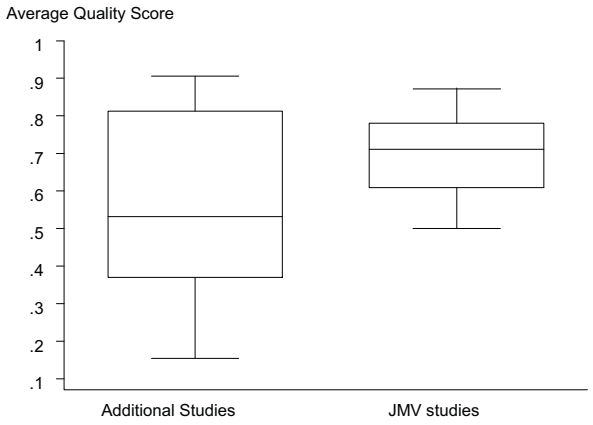

Figure: 2 Box plots of the Average quality score of studies (technology-centric studies only)

The box plots of the average scores for 26 studies found by the JMV study and the 16 additional studies are shown in Figure: 2. It is clear that some of the additional studies were of substantially poorer quality than studies found by the JMV study, but it is also clear that at least eight of the additional studies were of a comparable quality to those found by the JMV study. Thus, the quality evidence does not support the hypothesis that the JMV study intentionally rejected low quality papers.

3.4 Importance of the selected papers

We assessed the importance of the papers in terms of their citation index. We obtained citation indexes from SCOPUS (which included self-citations as well as citations in journal papers and conference papers) and Google Scholar (which included self-citations and citations in technical reports, power slides etc. as well as citations in journal papers and conference proceedings). The citation indexes are highly correlated (R^2 = 0.85, p<0.00001) with the Google Scholar value being much greater than the SCOPUS value (about 2.5 times greater). We report the results for SCOPUS in Figure 3. These results suggest that although some of the additional papers appear to be less important to the testing community, JMV do not seem to have had a policy of omitting less important papers.

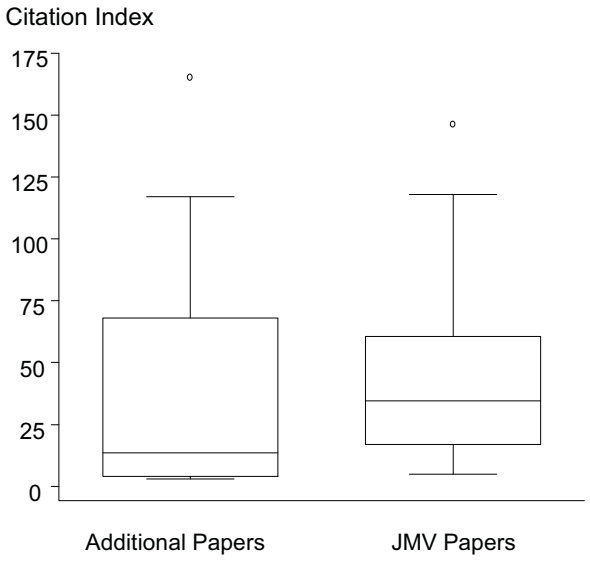

Figure 3: Citation Index (SCOPUS) of Additional Papers and JMV Papers (all papers)

3.5 The impact of the additional papers

Brereton and Kitchenham assessed the impact of the additional papers by extracting information about how the papers would be classified using the unit testing framework developed by JMV. They

extracted the information separately and then compared their results.

They looked for examples where the additional papers would either fill specific gaps in the framework (i.e. provide papers in a class that otherwise had no empirical papers), or identify missing elements in the framework (i.e. identify testing methods not included in the framework).

Only two of the papers missed by JMV would not have changed the JMV framework (Rothermel et al. 2001; Harrold et al., 2001). Rothermel et al. (2001) was an extended journal version of a conference paper found by JMV, it was also the paper with the largest SCOPUS citation index.

As shown in Table 5, the remaining eight papers would either have added a new element to the framework (3 papers), or added entries to an empty cell in the framework (5 papers). Three of the later papers all addressed specification testing.

Table 5: Impact of missing papers

Paper	Main classification	Impact on the JMV Framework
Netisopakul et al., 2002.	Test case generation/Test Model-based testing and Code-based testing.	Data coverage testing based on a test model not considered in framework.
Andrews and Zhang, 2003.	Test case generation/Formal Specification-based testing, Random testing, Functional testing.	Test cases based on formal specification included in framework but no example in JMV results.
Dupuy and Leveson, 2000.	Test set evaluation.	MCDC included in framework but no example.
Boyapati et al., 2002.	Test set generation/Specification-based (JML). Supported by the Korat tool.	Test cases based on formal specification included in framework but no examples.
Tan and Edwards, 2004.	Test case generation/Formal Specification-based. Supported by semi-automated tool.	Test cases based on formal specification included in framework but no example.
DeMillo and Offutt, 1993.	Test case generation/ Constraint-based testing supported by Godzilla tool Test case selection/Mutation-based.	Constraint based testing not included in JMV framework.
Untch, 1992.	Test set generation/Mutation-based testing/Program Schemata.	Mutation testing using schemas included in JMV framework but no examples.
Wappler and. Lammermann, 2005.	Test set generation/Evolutionary algorithms.	Testing using Evolutionary algorithms not included in the JMV framework.

4. Discussion

4.1 Proposition RQ1-P1

RQ1-P1 was formulated as follows:

> Where an SLR is conducted within well-defined parameters describing the topic, information sources and chronological bounds, the researchers should find the same set of sources.

Our results have shown that given the same basic research goal, with a search period and digital libraries clearly defined, two researchers can identify very different sets of primary studies. Furthermore their results had very little overlap in terms of identified primary studies with the previously published JMV study. Clearly the results are limited to novice researchers with experience neither in the research topic nor the SLR methodology. However, this result suggests that our previous study that found novices had very favorable experiences of using SLRs should be treated with some caution (Kitchenham et al., 2010b). That is, although the novice researchers reported that overall they enjoyed the experience, we cannot be sure how complete were the results of their SLRs. Furthermore, our results contrast starkly with results reported by MacDonnell et al. (2010) who found that two independent SLRs were very similar when produced by domain experts with experience of the SLR process, and addressing a well-defined and very specific research question. It may be that our study and MacDonnell et al.'s study represent the extremes of repeatability with our case representing the worse performance of the SLR methodology and MacDonnell et al. reporting the best.

Although the RAs started with different research questions which influenced the papers found by their search process, it was also clear that applying inclusion/exclusion criteria was a significant problem. The RAs included 19 papers that were rejected by more experienced researchers. This suggests that repeatability depends not only on using the same

inclusion/exclusion criteria but employing them in the same way.

After applying the inclusion/exclusion criteria, Brereton and Kitchenhan found 10 papers that were not included in the JMV study. Furthermore, we could not find any reason related to the quality or importance of the papers to explain why at least some of the 10 papers should have been omitted. The JMV study did not claim to have used the SLR methodology they only claimed to have done an "extensive search" which they must have regarded as appropriate for use in the context of evidence-based software engineering. Thus, the results confirm potential problems with the rigour and completeness of literature reviews that *do not* use a well-specified SLR process.

4.2 Proposition RQ1-P2

RQ1-P2 was formulated as:

> Where an SLR is repeated using the same parameters and guidelines for analysis, then the analyses will produce the same conclusions.

Since our case studies found very different sets of primary studies, we were not able to address this question directly. However, we were able to assess the impact of missing studies for the JMV results. This issue relates to the impact on conclusion repeatability between different literature reviews.

We found that 8 of the 10 additional papers that we believe addressed the JMV research questions, would have changed the JMV results in terms of adding three new classes to the testing framework and populating three otherwise empty classes. We can conclude that missing papers can have a substantial impact on conclusions. This contrasts with our previous case study (Kitchenham et al., 2010a), where we found little change in the results of two SLRs addressing the same research question, one of which used a restricted manual search and the other that used a broad automated search. However, the difference may be because the Kitchenham et al. (2010a) study was a mapping study and the publication trends did not change much.

4.3 Limitations

This study suffers from the normal problem with case studies that it is difficult to generalize the results. We cannot be sure whether the results we obtained were due to the specific RAs or would be the same for any pair of novice researchers.

In addition, the RAs developed their own research questions from a general research topic. Thus, the differences between their results might have been due to differences in the research question rather than to other aspects of the SLR search and selection process. The reason we allowed the RAs to develop their own questions is because the specification of the research question is considered to be the first part of the SLR process. However, with the benefit of hindsight, it would have been better to have given the RAs exactly the same research question rather than the same general topic.

It must also be emphasized that the JMV study did not make any claim to be an SLR and did not report their primary selection process in any detail. Thus, our conclusion that missing papers can have an influence on literature review conclusions refers to expert literature reviews not SLRs.

5. Conclusion and Future Work

Our results indicate that having the same broad research topic will not guarantee repeatability with respect to the identification and selection of primary studies if those studies are undertaken by novices. In addition, inclusion/exclusion criteria present difficulties for novices. However, it is also clear that expert literature reviews that do not use a rigorous search and selection process may miss papers that would have a serious impact on the results of the review. The best hope for repeatability seems to be to use researchers with expertise in the SLR process and the topic being reviewed. Furthermore, researchers undertaking literature reviews cannot expect their results to be repeatable unless they fully document their search process (including any search strings) and also their inclusion/exclusion criteria.

The JMV study not only appeared to have missed relevant papers in the digital libraries they searched, they also restricted their search to the IEEE and ACM digital libraries. We are currently undertaking an SLR of unit testing and regression testing using the JMV study as a baseline to construct search strings. This aims to look for more missing papers by searching more sources using a search process that can (to a certain extent) be validated for completeness.

References

Andrews, J.H. and Zhang, Y. 2003. General test result checking with log file analysis, IEEE TSE, 29 (7), pp. 634-648.

Boyapati, C., Khurshid, S. and Marinov, D. 2002. Korat: automated testing based on Java predicates, Proceedings of the 2002 ACM SIGSOFT

International Symposium on Software Testing and Analysis, pp 123-133.
Budgen, D. 2009. Supporting Novices undertaking Systematic Literature Reviews: EPIC Case Study Protocol No: CS001/07, Technical Report edn, Durham and Keele Universities.
DeMillo, R. A. and Offutt, A. J. 1993. Experimental results from an automatic test case generator, TOSEM, 2(2), pp 109-127.
Dupuy, A. and Leveson, N. 2000. An empirical evaluation of the MC/DC coverage criterion on the HETE-2 satellite software, Proceedings of the 19th Digital Avionics Systems Conferences, volume 1, pp 1B6/1-1B6/7.
Frankl, P.G., Weiss, S.N. and Hu, C. 1997. All-Uses versus Mutation Testing: An Experimental Comparison of Effectiveness, JSS, 38, pp. 235-253.
Harrold, M.J., Jones, J. A., Li, T., Liang, D., Orso, A., Pennings, M., Sinha, S., Spoon, S. A. and Gujarathi, A. 2001. Regression test selection for Java software, ACM SIGPLAN Notices, 36(11).
Jørgensen, P. 2008. Software Testing: A Craftman's Approach, 3rd edn, Auerbach Publications.
Juristo, N., Moreno, A.M., Vegas, S. and Solari, M. 2006. In Search of What We Experimentally Know about Unit Testing, IEEE Software, vol. 23, pp. 72-80.
Juristo, N., Moreno, A.M. and Vegas, S. 2004. Reviewing 25 Years of Testing Technique Experiments, Empirical Software Engineering, 9(1), pp. 7-44.
Kitchenham, B.A. and Charters, S. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE-2007-01.
Kitchenham. B.A., Burn, A.J. and Zhi, L. 2009. A Quality Cheklist for Technology-Centred Testing Studies, Proceedings 13th International Conference on Evaluation and Assessment in Software Engineering (EASE), BSC eWic.
Kitchenham, B., Brereton, P., and Budgen, D. 2010a. The Educational Value of Mapping Studies of Software Engineering Literature. ICSE 2010 Education Theme. South Africa, ACM Press, pp 1-7.
Kitchenham, B.A., Brereton, O.P., Turner, M., Niazi, N., Linkman, S., Pretorius, R. and Budgen, D. 2010b. Refining the systematic literature process - two participant-observer case studies. ESJ, 15(6), pp 618-653.
MacDonnell, S., Shepperd, M., Kitchenham, B. and Mendes, E. 2010. How Reliable are Systematic Review in Software Engineering. IEEE TSE 36(5), pp 676-687.
Netisopakul, P., White, L. J. and Morris, J. 2002. Data coverage testing, The 9th Asia-Pacific Software Engineering Conference, pages 465-472.
Roper, M., Wood, M. and Miller, J. 1997. An empirical evaluation of defect detection techniques, Information and Software Technology, 39(11), pp. 763-775.
Rothermel, G., Untch, R. H., Chu, C. and Harrold, M. J. 2001. Prioritizing test cases for regression testing, IEEE TSE 27(10), pp 929-948.
Tan, R.P. and Edwards, S. H. 2004. Experiences evaluating the effectiveness of JML-JUnit testing, ACM SIGSOFT Software Engineering Notes, 29(5), pp 1-4.
Untch, R. H. 1992. Mutation-based software testing using program schemata, Proceedings of the 30th Annual Southeast Regional Conference, pp 285-291, Raleigh, North Carolina, USA
Wappler, S. and Lammermann, F. 2005. Using evolutionary algorithms for the unit testing of object-oriented software, Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, pp 1053-1060, Washington DC, USA
Wong, W.E., and Mathur, A.P. 1995. Fault Detection Effectiveness of Mutation and Data-flow Testing, *J* Software Quality Journal, 4(1), pp 69-83.
Yin, R., K. 2003. Case Study Research: Design and Methods, 3rd edn, Sage Publications Inc.

Acknowledgements

We thank the members of the EPIC project, particularly, Feng Bian, for their contributions to this study. The work described in this paper was funded by the UK Engineering and Physical Sciences Research Council project EPIC/E046983/1.