Investigating bias in the search phase of Software Engineering secondary studies

José Amancio Macedo Santos[1], Alcemir Rodrigues Santos[2], and Manoel Gomes de Mendonça[3]

[1] State University of Feira de Santana, Technology Department, Bahia, Brazil
zeamancio@ecomp.uefs.br,
[2] Reuse in Software Engineering Lab – RiSELabs
Federal University of Bahia, Bahia, Brazil,
alcemirsantos@dcc.ufba.br,
[3] Fraunhofer Project Center for Software & Systems Engineering,
Federal University of Bahia, Bahia, Brazil,
mgmendonca@dcc.ufba.br

Abstract. *Context.* Researchers are increasingly resorting of secondary studies (*e.g.* systematic literature reviews and mapping studies) in Software Engineering. This method is strongly dependent on the source of primary studies adopted, which is a bias. We did not find guidelines or benchmarks to evaluate the sources in a systematic way. *Objective.* In this paper we aim to tackle the selection of electronic data sources while conducting such kind of studies evaluating the equilibrium between the volume and number of relevant papers. *Method.* In this sense, we proceed towards a secondary study to analyze the overlapping of three different electronic data sources. We also compared our results with other similar studies. *Results.* Our results show minimum overlapping and no effortless combination of electronic data sources at all. *Conclusion.* We conclude that researchers shall resort of completeness to work with a feasible set of papers to review. Specially in secondary studies adopting general and no standardized terms.

1 Introduction

Systematic secondary studies (SS), such as systematic literature review (SLR) or mapping study (MS) [13, 23] have been widely adopted in Software Engineering. They are based on search and synthesis of primary studies. In this work, we focus on the search phase where the rigor is one factor that distinguishes SS from traditional reviews. In such phase, the search can be automatic (by using digital libraries and search engines) or manual (by looking into specific journals and/or proceedings). The definition of the search strategy must consider the equilibrium between the large volume of papers on the available sources and the number of relevant primary studies. Kitchenham declared that digital libraries are insufficient for a full SS [13]. Later, on the other hand, Silva *et al.* also pointed evidence of limitations in the manual research process [21].

In this paper, we focus on automatic search considering different aspects, which we believe to be essential to achieve the best result from such a research process. First, the *digital resource database choice*, which includes both, digital libraries and search engines. From now on, we adopt Electronic Data Source (EDS) to refer either to digital libraries or search engines indistinctly. Currently, it is common the use of different digital libraries (*e.g.*, `IEEEXplore`, `ACM Digital Library`, and `ScienceDirect`) or different search engines (*e.g.*, `Web of Science`, `Scopus`, and `Google Scholar`), in the search for primary studies in a SS. Such large number of EDS may be explained by the heterogeneity of primary study published by each of the digital libraries.

The problem of the heterogeneity is insufficiently explored and few works addressed this issue. To the best of our knowledge, only Chen *et al.* compared papers from different EDS and proposed metrics to characterize them [3]. They declared that rather than a systematic EDS choice, "researchers select them mainly based on personal knowledge, experience, preferences and/or recommendations by peers". In fact, it does exists literature comparing EDS from a perspective other than their analyze as primary sources for SS. Achambault *et al.* compared two search engines, but from the perspective of the bibliometric statistics [1]. Other two works [16, 17] addressed differences on citations from two different search engines.

Differences on the amount of primary studies retrieved by each EDS become more evident for search strings using either generic or unstandardized terms. In such cases, the large volume of papers makes very difficult to combine EDS in the search phase. The alternatives are to perform a manual search, to choose only one out of the available EDS, to restrict the search string, or combine some of these options. For instance, Zhang *et al.* [26] limited their search string due to the huge search space while looking for papers addressing design problems, *a.k.a.* "code smell". Sjøberg *et al.* [22] mapped the scene of "controlled experiments" adopting a manual approach as search strategy and their set of primary studies was the basis to four SLR [6, 7, 11, 12]. In addition, Dieste *et al.* [5] studied the problem of generality and absence of standardization for controlled experiments. They analyzed the optimality of a number of search strategies for SS. We will discuss other cases later.

In this paper, we investigate the low precision problem of the EDS search based strategy for studies with generic terms. We aim to produce empirical evidence about the existing differences among the EDS and perhaps motivate novel discussions on the topic. We address the issue by performing a MS based on generic terms and focused our analysis in the primary studies selection phase. Our main finding strengthens that the combination of EDS in SS adopting such generic terms needs too high effort, which is normally impractical due research groups limitations of time and human resources. Another finding is related to the definition of search strategies. Based on the gathered evidences, it is possible to state that the currently SS presented has been biased, which consequently affect their findings. Finally, we noticed the need of empirical studies to better evaluate and to define the suitable search strategies in a systematic way. A

transversal contribution of this work is the presentation of the process that we adopted in the selection of primary studies. We performed the activity with 57 graduate students (Master and Ph.D.), such large amount of human resources is uncommon and may support in the selection phase of that kind of broad studies.

The rest of the paper is structured as follows. Section 2 summarizes the settings of our study including the research questions and characteristics of the EDS adopted. Section 3 describes the data collection process. Sections 4 and 5 present the results and discuss them. Section 6 discusses the limitations of the study. Section 7 summarizes related work, which tackled similar topics. Finally, Section 8 presents our conclusions and enumerates some future work.

2 Study Settings

2.1 Research Questions

It is necessary a proper guidance to achieve the objectives of any piece of research. In this sense, we defined three research questions to address the EDS low precision problem, which we enumerate next. In time, we adopted three well known EDS to perform our analysis: `IEEEXplore`, `ACM Digital Library`, and `Scopus`. In addition, we rely on the Kitchenham *et al.* [14, 15] work to compare the number of relevant papers among different SS. The following research questions guide our investigation.

RQ1: What is the overlap among the papers retrieved by `IEEEXplore`, `ACM Digital Library`, and `Scopus` EDS?
RQ2: Is it possible to combine `IEEEXplore`/`ACM Digital Library`/`Scopus` to reach better results?
RQ3: Is there any pattern in terms of the amount of relevant papers among different secondary studies?

2.2 Electronic Data Sources

In this study we used three EDS: two publishers (`IEEEXplore` and `ACM Digital Library`) and one indexer (`Scopus`). `IEEEXplore` (IEEEX) is a powerful online resource for accessing scientific and technical publications produced by the Institute of Electrical and Electronics Engineers and its publishing partners. `ACM Digital Library` (ACMDL) is also an online resource that serves ACM members and the computing profession with leading-edge publications, conferences, and career resources. According Scopus' host (Elsevier) homepage, `Scopus` is the largest abstract and citation database of peer-reviewed literature.

3 Data Collection Process

In this section, we describe how we carried this investigation, including the strategy used to define the search string, the selecting of the relevant papers, and the process adopted to select other similar SS and compare the results.

3.1 Search String Definition

We adopted the string defined by Dieste and Padua [5], because of the similarity with our study. They adopted `precision` and `sensitivity` measures in the evaluation of the strings. While sensitivity measures the fraction relevant material (*i.e.*, papers) retrieved, precision measures the fraction of the material retrieved that is relevant. We adopted the string with higher precision value, because we noticed that the use of the word "experiment" leads to a huge amount of data and perhaps it could turn our work even harder. Table 1 shows the main strings presented by Dieste and Padua [5]. We adopted the string of the first line.

Table 1. Initial search string [5].

Search String	**S**	**P**
experiment OR experimental study OR experimental analysis OR experimental evidence OR experimental setting	83.3%	20.7%
experiment OR empirical study OR experimental study OR empirical evaluation OR experimentation OR experimental comparison OR experimental analysis OR experimental evidence OR experimental setting	93.3%	17.6%

S: Sensitivity; **P:** Precision

Table 2 shows the numbers of papers retrieved in our searching. First, we show results using exactly the same string proposed by Dieste and Padua [5], considering the range of year lasting from 2003 to 2012. As we can see, the number of papers retrieved was very high. Therefore, we tried two other strategies. First, we restricted to papers only on the field of "Computer Science". Also in this case, the number of papers made impractical the realization of our study. Finally, we introduced the term "Software Engineering" and limited our searching to the year of 2011. By convenience, as we started this study on the year of 2012 we decided to investigate the problem for the year of 2011.

3.2 The Selection Process of Controlled Experiments

Human Resources. The selection of papers was made up by graduate students enrolled in the Experimental Software Engineering course, of the Computer Science program at Federal University of Bahia. We defined the protocol, controlled the distributions of papers and supervised the process. The students receive grades for their participation in the study.

We applied the activity with students from two classes: (i) the second term of 2012 (2012.2) and (ii) the first one of the year 2013 (2013.1). In total, 57 students worked on the data collection, 28 from 2012.2 class and 29 from 2013.1 class. The students had to analyze the papers recovered by the search with

Table 2. Initial results of the automated search.

EDS	Years	Search String		
		[5]	*CS*	*SE*
Scopus	2003-2012	2.820.467	277.281	16.120
	2011	.	.	2.410
IEEEX	2003-2012	432.372	295.335	15.807
	2011	.	.	2.177
ACMDL	2011	.	.	1.093

CS: Only Computer Science; SE: Added "Software Engineering"

the string previously defined. Then, they identified relevant papers according to inclusion/exclusion criteria, which was the same presented by Sjøberg et al. [22]. The 2012.2 term students analyzed 2410 papers from the Scopus library. In addition, the 2013.1 term students analyze 1093 papers from the ACMDL, as well as, 2177 papers from the IEEEX.

Training. The students were trained during the course. Besides systematic literature reviews and mapping studies, the Experimental Software Engineering course discussed controlled experiments deeply. In addition, we asked to the students to read important papers in this area, such as the Sjøberg et al. [22] and Jedlitschka et al. [9]. Such reading provides a broad comprehension of controlled experiments.

During the training, we presented the selection process. Figure 1 illustrates our process. We used the *search string* to build an *Initial List of Papers*, then together we (researchers and the students trained) processed this list of papers in three subsequent phases: (i) the *Studies Selection Phase*; (ii) the *Consensus Phase*; and (iii) the *Data Extraction Phase*). In the end of the process we transformed the initial list papers in two lists: (i) one with the rejected papers; and (ii) other with the accepted papers (the relevant papers for our study).

Studies Selection Phase. After the training we presented the protocol of our study. The protocol was based on the mapping study presented by Sjøberg et al. [22]. Then, the students received a spreadsheet with basic information of each paper retrieved from EDS, including id, title, abstract, keywords, authors, number of pages and publishers.

Two students revised each paper simultaneously and independently. By using basic combination, we arranged all possible pairs of students and randomly assigned pairs of students to papers. Then, we sent a list of papers designed for each student, individually. This increased the difficult of a student find the others reviewers of the paper assigned to him/her.

The first task assigned to the students was judge the list of papers according to the study protocol. They had to mark each paper with accept, reject or doubt. In the case of rejection, they had to pinpoint which exclusion criterion was adopted. Afterwards, each student individually returned his/her list of papers with the judgment attached.

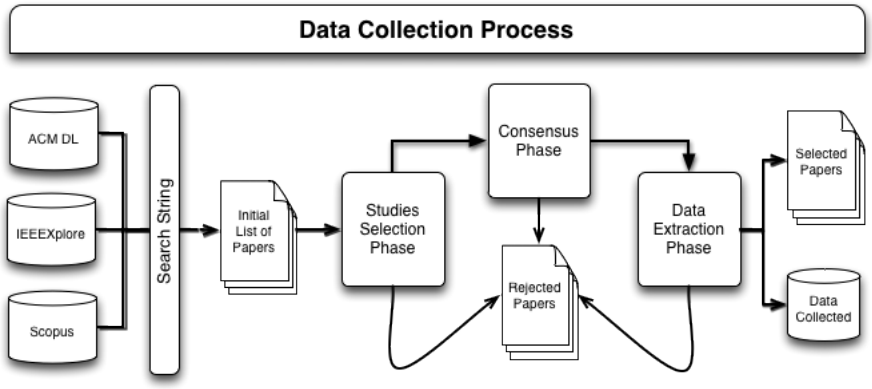

Fig. 1. Data Collection Process

After gathering the students' judgment, we processed the answers with the following criteria. Papers judged by both reviewers as "accepted" were considered as a relevant paper. In this case, the paper follows to the "Extraction Phase". On the other hand, papers judged by both reviewers as "rejected" were disregarded to the next phase. Furthermore, whatever other combination of revisions were considered to following phase, before the "Extraction Phase". This intermediate phase we called "Consensus Phase".

Consensus Phase. We used this phase to discuss the papers judged as "doubt" in the "Studies Selection Phase" and papers that produced disagreement among reviewers. We did this randomly rearranging the students in pairs. The papers were also randomly redistributed. We carried the paper redistribution independently from the initial distribution in the selection phase. Thus, each student pair received a new list of papers to review. Differently from the previous phase, this time the student pair worked together to solve doubts and disagreements from the previous reviewers.

We performed the consensus phase in a lab. Each pair analyzed the papers considered to this phase with the supervision and support of one researcher signing this work. The supervision helped us to mitigate new doubts and to make sure only papers of our interest were judged "accepted". Only in case of insufficiency the basic information to solve the doubts the papers were downloaded and read. After this phase, we started the "Data Extraction Phase".

Data Extraction Phase. In this phase we downloaded all the papers judged as "accepted" in the previous phases. We repeated the process of arrangement and papers redistribution to the pairs. During this phase, after a detailed reading, the students detected some papers out of the scope of this study and asked one researcher to reject them through a mailing list. The researchers defined about the paper rejection. This phase helped to increase the confidence on the set of

relevant papers. We do not detail the extraction phase because is out of scope of this work.

3.3 Secondary Studies Selection

We compared our results with other SS. We rely this comparison on several studies found on two tertiary studies carried by Kitechenham *et.al* [14][15]. While the first one [14] points out 35 SS, the second one [15] points out 20. From these 55 studies we selected those using general terms in the topics addressed in the papers. At first, we used titles and abstracts to decide whether the studies should be used or not to extract information about search strategies and selecion phase. Then, we downloaded the candidate SS to read and decide if they had similar characteristics of our study, *i.e.*, if they are broad and with general terms. Additionally, the authors discussed such study selection to avoid irrelevant papers in the comparison. In the end, we extracted data from 8 out of 55 initial SS and from the tertiary studies [14] and [15].

4 Results

In this section, we present the results of our study, considering each of our research questions.

4.1 EDS Results Overlapping

Our first research question addressed the overlap among the papers retrieved by Scopus, IEEEXplore and ACM DL. Is there an overlap among them? In other words, how many papers published in one EDS were also available in the others? Figure 2 shows the answer, which we discuss next.

First, by observing the number of papers retrieved by each of the EDS we found that Scopus and IEEEX had similar values (2410 and 2177, respectively), but ACMDL had a smaller amount of papers (1093) (Table 2). Afterwards, we analyzed the uniqueness and the overlapping of papers along the results of each EDS. Figure 2(a) shows their overlap, which is characterized as follows: 135 papers appeared in the ACMDL and Scopus; 35 appeared in ACMDL and IEEEX; 249 appeared Scopus and IEEEX; and 11 papers appeared in ACMDL, IEEEX, and Scopus. Therefore, 5239 (912+35+1882+135+11+249+2015) out of 5680 (2410+2177+1093) were unique papers.

These numbers show that the overlap on the raw results of the EDS was low. If we consider the total papers that we found, the number of common papers was insignificant, since 11 papers represent only *0.21%* of the total (5239). Additionally, the best relation found takes place between IEEEX and Scopus. Considering the total number of papers that we found in both EDS, there were 260 (11+249) out of 4327 (35+1882+135+11+249+2015), representing just *6.0%* of overlapping.

We also analyzed the uniqueness and the overlapping after the studies selection phase (Figure 2(b)). In that phase, we selected 15 papers from ACMDL, representing *1.4%* of its 1093; 46 papers from IEEEX, representing *2.1%* of its 2177; and 59 papers from Scopus, representing *2.4%* of its 2410. There were 2 papers simultaneously in ACMDL and Scopus, which corresponds to *2.8%* of the total of 72 (13+24+53) papers and 4 papers simultaneously in IEEEX and Scopus, which corresponds to *4.0%* of the total of 101 (42+24+53) papers. After selection of relevant papers, the total of unique papers was 114 (*2.2%* of 5239).

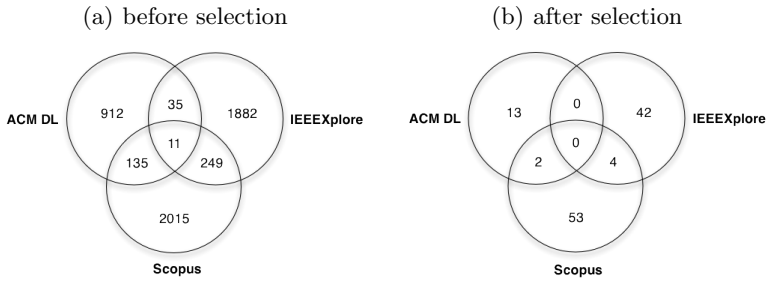

Fig. 2. Number of papers and intersections among EDS.

4.2 Search Strategies Relying on EDS Combination

Next research question addresses the strategies that aim at maximization of the number of papers reached while carrying a SLR. In other words, is it possible to optimize the results by combining EDS? First of all, we calculated values of sensitivity and precision for each combination among the three EDS. We consider our universe set ($|U| = 5239$), the number of relevant papers ($|R| = 114$) and the number of irrelevant papers ($|I| = 5125$). For this analysis, we disregard the relevant papers that our string was unable to retrieve, *i.e.*, we are considering to have found all relevant papers.

Table 3 shows the values of precision and sensitivity for each EDS, as well as the values regarding its possible combinations. The values of sensitivity increase as soon as we combine the EDS. For instance, ACMDL itself reached 13.2% of sensitivity (15 papers out of the 114 relevant) and its combination with IEEEX reached 53.5% (61 papers out of the 114 relevant). This happens for all combinations, which was predictable, since we detected a lack of overlap among the search machines when addressing the RQ1.

Dieste and Padua [5] inferred four scales to evaluate search strategy from previous studies carried in the medicine field (Table 4). We adopt their scale as parameters to evaluate the quality of search strategies. They believe that researchers should aim to reach the optimum strategy thresholds or, at least,

Table 3. Values of sensitivity and precision

Search Machines	Sensitivity		Precision	
	Values	%	Values	%
ACMDL	15/114	13.2%	15/1093	0.3%
IEEEX	46/114	40.4%	46/2177	0.9%
Scopus	59/114	51.8%	59/2410	1.1%
ACMDL ∪ IEEEX	61/114	53.5%	61/3224	1.2%
ACMDL ∪ Scopus	72/114	63.2%	72/3357	1.4%
IEEEX ∪ Scopus	101/114	88.6%	101/4129	1.9%
ACM ∪ IEEEX ∪ Scopus	114/114	100%	114/5239	2.2%

the acceptable strategy values. Our results did not reach any of those thresholds. This fact might indicates that we cannot define an optimal search strategy based on EDS to such kind of secondary studies.

Table 4. Search Strategy scales [5].

Strategy type	SR	PR	Goal
High Sensitivity	85-99%	7-15%	Maximum sensitivity despite poor precision
High Precision	40-58%	25-60%	Maximum precision despite poor sensitivity
Optimum	80-99%	20-25%	Maximization of both ranges (sensitivity and precision)
Acceptable	72-80%	15-25%	Good enough sensitivity and precision

SR: Sensitivity rabge; PR: Precision range.

4.3 Comparing Results on Secondary Studies

The last research question addresses if there is any pattern in terms of the amount of selected papers among different SS. In other words, we looked for similarities among our study and the results achieved by other studies. We extracted a set of secondary studies using general terms and covering a wide range of years from Kitchenham's works [14] [15]. Table 5 shows the chosen related studies.

Table 6 shows the range of years and the total amount and relevant papers of each SS we chose to analyze. Only one [15] out of ten reviews adopted an automatic searching. Two ([18] and [19]) adopted a mixed strategy by using both, manual and automated approaches. Another four adopted only manual search [14][20][8][25], and three studies lack on discuss its strategy of research.

Table 5. Studies chosen, based on generic terms or widely (in years)

Reference	Aim
[14]	Assess the impact of systematic literature reviews in EBSE
[18]	Investigate the rigor of claims arising from Web engineering research
[20]	Examine the state of computer science research
[25]	Analyzes quantity and quality of empirical evaluations
[10]	Analyze the maturity level of the knowledge about testing techniques by examining existing empirical studies about these techniques
[2]	A review paper on the software inspection process. It also examines experimental studies and their findings
[15]	Provide an annotated catalogue of SLRs available to soft- ware engineering researchers and practitioners
[4]	Examines the maturation of the software architecture research area
[8]	Provide an assessment of the status of empirical software research
[19]	Assess the effects of software reuse in industrial contexts

In these last three studies, they present theoretical discussion about an specific area and perhaps the authors searched for the most important and well known papers of those areas.

Table 6. Studies based on generic terms or widely (in years) to identify a mapping of an area

Reference	Strategy	Period	Results	PR
[14]	Manual	2004 to 2007(4 years)	2506	20
[18]	Mixed	No more than 2 years for journals and proceedings, and 9 years for IEEEX and ACMDL	unclear	173
[20]	Manual	1995 to 1999 (5 years)	By sampling: 628	628
[25]	Manual	1975 to 2005 (29 years)	1227	63 (5%)
[10]	Ad-Hoc	25 years	unclear	unclear
[2]	Ad-Hoc	25 years	unclear	unclear
[15]	Automated	2004 to 2008 (5 years)	1757	35
[4]	Ad-Hoc	1985 to 2006	Unclear	750
[8]	Manual	1996 to 2006	133	133
[19]	Mixed	1994 to 2005	unclear	11

Regarding the number of selected papers, only Zannier [25] presented the fraction of relevant papers in the amount of selected papers (5%). Only Kitchen-

ham *et.al* [14] [15] presented the amount of primary studies selected (0.8% and 1.9%, respectively). In our case, primary studies represent 2.2% from the papers retrieved by the EDS. For all cases, primary studies represent less than 5%.

5 Discussion

The first observation we done was related to differences on results for each EDS. We found best measures for Scopus, than ACMDL and IEEEX. This can be explained by the differences between the use of publishers and indexers. Scopus is an indexer subscribing other different sources. Due this, it more likely this kind of EDS found more relevant (and irrelevant) papers. An aspect that might be considered is the relationship between the relevant and irrelevant new papers. This is also dependent of the quality of the papers and it is another confounding variable to be considered in SS.

Our empirical results also evidence one of the aspects that make a SS so hardly work, specially for cases where terms are generics or non standardized. To get more embracing results it is necessary to combine different EDS. This seems prohibitive. It is almost impractical to combine a significant number of EDS in opposite to a limited search based on small number of EDS or a manual strategy. For us, this indicates that, like in other experimental methods, SS related to similar topics are necessary to guarantee more confident conclusions. This is not very usual. Wohlin et al. [24] address this subject, comparing two mappings studies on the same topic.

Another important finding is that the current SS have been strongly biased in their search strategy. One evidence is the low overlapping among EDS. The Figure 2 shows that there was not papers subscribed by the three EDS adopted in our study, after the selection phase. This can be expected in many other SS, because ACMDL and IEEEX are publishers, not indexers. Another evidence of bias in SS is the low precision values for each individual EDS added to the lack of studies that combine high number of EDS. We believe that more empirical studies are necessary to produce a benchmark of sensitivity and precision on the EDS combination. This would help researchers in the definition of their search strategy and would help to mitigate the bias of search strategies in secondary studies.

Finally, the analysis of the other SS (as presented in Section 4.3) reinforced our belief that the problem of bias on search strategies in SS needs to be discussed. The small number of precision obtained by other SS is in according to our study. We consider that this can motivates other discussions on the topic.

6 Limitations

We detach three limitations of our study. The first one is the adopted selection method to identify relevant papers. We presented the method in Section 3.2. It was unusual method. The selection was performed with two groups in two semesters and small differences on the process occurred. Some aspects mitigate

this limitation. The process was evaluated by the course professor and we had long time to select relevant papers independently by each participant, with a supervised consensus phase in a lab, including extracted data from the selected papers, which we did not discuss in this paper because is out of scope.

Another limitation is related to the use of an extensive number of graduate students. Their motivation was not related to the research, but to the course work. However, we highlight that many graduate students have to perform SLRs, therefore, it is true they hold interest in whole process.

We also have to consider the search string. We defined our search string based on the Dieste et al.[5] study and added the term "Software Engineering". Due to problem of generalization we do not know how much near from the ideal relevant papers we are. But, despite this, we believe that because we are based on another study, our string search is in accordance with the controlled experiment topic.

Lastly, we did not perform a tertiary study identifying a set of SS in Software Engineering. We based on two other studies, with limited search strategies. However, we believe that these studies are relevant because their topics were very specific instead of general; and the authors are highly referenced in Experimental Software Engineering. Despite the subjectivity of our tertiary studies selection criteria the authors discussed the choices to mitigate threat to validity.

7 Related Works

Sjoberg *et al.* [22] reported upon the state of how controlled experiments in software engineering were conducted and the extent to which relevant information was reported. They selected `103` papers out of those published in 12 leading software engineering journals and conferences in the decade from 1993 to 2002 (`5,453`). The selected papers report controlled experiments in which individuals or teams performed one or more software engineering tasks. This work highly influenced later works, such as *(i)* Dieste and Pádua [5], which used Sjoberb's work as a gold standard, *(ii)* and four systematic reviews based on the same set of controlled experiments [11][12][7][6].

The Dieste and Pádua [5]'s work is close to ours, in the sense that they analyze the optimality of search strategies for use in systematic reviews. From different combinations of terms, they evaluated `sensitivity` and `precision` of several search strategies aiming to find an optimum strategy. They identified trends and weaknesses in terminology used in articles reporting experiments.

Chen et.al [3] compared papers found in different EDS and proposed metrics to characterize them. They proposed an initial set of metrics for characterizing the EDS from the perspective of the needs of secondary studies. Other works compare EDS, but not from the perspective of primary sources for secondary studies. For instance, Achambault *et.al* [1] compared two EDS from the perspective of the bibliometric statistics and Meho and his colleagues [16][17] investigate citation counting and ranking, and h-index of human-computer interaction researchers and impact of information studies based on Scopus and Web of Science.

8 Conclusion

In this paper we investigated the population selection problem through conducting a SS on controlled experiments in software engineering. We took this path in order to carry a quantitative analysis on the overlapping of papers from three different electronic sources: `IEEEX`; `ACMDL`; and `Scopus`. Our results showed minimum overlapping and no effortless combination of EDS at all to conduct such kind of secondary study.

We concluded that researchers should resort of completeness to work with a feasible set of papers to review. Specially, in secondary studies searching literature by containing general and no standardized terms, such as "controlled experiments". In other words, the presence of general terms in the search string lead search matching to unfeasible sets of papers to review. For instance, with the only three electronic sources that we used in this study, we reach `5239` unique papers published in the year of `2011` that might contain relevant content to our research. In fact, only `2.2%` of them do matter to us.

In addition, we compared the percentage of relevant papers selected of our findings with other SS. We concluded that such values seemed to be a trend on such kind of studies. Therefore, we advocate the need of different secondary studies in the same research field to accomplish a broad view of the literature.

Unfortunately, additional work is needed regarding the secondary studies population selection problem. In future work, we may work towards to create a benchmark of thresholds, it would allow researchers to compare his secondary studies with different others previously carried.

Acknowledgments: *We would like to thank Cleber Pereira dos Santos and the participants of the Experimental Software Engineering courses. This work was partially supported by FAPESB, SECTI-BA (Bureau of Science and Technology of Bahia), and Fraunhofer Project Center for Software & Systems Eng. agreement 2012/001.*

References

1. Archambault, E., Campbell, D., Gingras, Y., Larivière, V.: Comparing bibliometric statistics obtained from the web of science and scopus. Journal of the American Society for Information Science and Technology 60(7), 1320–1326 (2009)
2. Aurum, A., Petersson, H., Wohlin, C.: State-of-the-art: software inspections after 25 years. Software Testing, Verification and Reliability 12(3), 133–154 (2002)
3. Chen, L., Babar, M.A., Zhang, H.: Towards an evidence-based understanding of electronic data sources. In: Proc. of the 14th EASE. pp. 135–138 (2010)
4. Clements, P., Shaw, M.: The golden age of software architecture: A comprehensive survey. Tech. rep. (2006)
5. Dieste, O., Padua, A.G.: Developing search strategies for detecting relevant experiments for systematic reviews. In: Proc. of the 1th ESEM. pp. 215–224 (2007)
6. Dybå, T., Kampenes, V.B., Sjøberg, D.I.: A systematic review of statistical power in software engineering experiments. Information and Software Technology 48(8), 745 – 755 (2006)
7. Hannay, J.E., Sjoberg, D.I.K., Dyba, T.: A systematic review of theory use in software engineering experiments. IEEE Trans. Softw. Eng. 33(2), 87–107 (2007)

8. Höfer, A., Tichy, W.F.: Status of empirical research in software engineering. In: Empirical Software Engineering Issues. Critical Assessment and Future Directions, LNCS, vol. 4336, pp. 10–19 (2007)
9. Jedlitschka, A., Ciolkowski, M., Pfahl, D.: Reporting experiments in software engineering. In: Shull, F., Singer, J., Sjberg, D.I.K. (eds.) Guide to Advanced Empirical Software Engineering, pp. 201–228. Springer London (2008)
10. Juristo, N., Moreno, A.M., Vegas, S.: Reviewing 25 years of testing technique experiments. Empirical Softw. Eng. 9(1-2), 7–44 (2004)
11. Kampenes, V.B., Dybå, T., Hannay, J.E., K. Sjøberg, D.I.: A systematic review of quasi-experiments in software engineering. Inf. Softw. Technol. 51(1), 71–82 (2009)
12. Kampenes, V.B., Dybå, T., Hannay, J.E., Sjøberg, D.I.: A systematic review of effect size in software engineering experiments. Information and Software Technology 49(1112), 1073 – 1086 (2007)
13. Kitchenham, B.: Guidelines for performing systematic literature reviews in software engineering. Technical Report Version 2.3, Keele University and University of Durham, UK (2007)
14. Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., Linkman, S.: Systematic literature reviews in software engineering - a systematic literature review. Inf. Softw. Technol. 51(1), 7–15 (2009)
15. Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., Linkman, S.: Systematic literature reviews in software engineering - a tertiary study. Inf. Softw. Technol. 52(8), 792–805 (2010)
16. Meho, L.I., Rogers, Y.: Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of scopus and web of science. J. Am. Soc. Inf. Sci. Technol. 59(11), 1711–1726 (2008)
17. Meho, L.I., Sugimoto, C.R.: Assessing the scholarly impact of information studies: A tale of two citation databases– scopus and web of science. J. Am. Soc. Inf. Sci. Technol. 60(12), 2499–2508 (2009)
18. Mendes, E.: A systematic review of web engineering research. In: Proc. of ISESE. pp. 509–518 (2005)
19. Mohagheghi, P., Conradi, R.: Quality, productivity and economic benefits of software reuse: A review of industrial studies. Empirical Softw. Engg. 12(5), 471–516 (2007)
20. Ramesh, V., Glass, R.L., Vessey, I.: Research in computer science: an empirical study. Journal of Systems and Software 70(12), 165 – 176 (2004)
21. da Silva, F.Q.B., Santos, A.L.M., Soares, S., França, A.C.C., Monteiro, C.V.F., Maciel, F.F.: Six years of systematic literature reviews in software engineering: An updated tertiary study. Inf. Softw. Technol. 53(9), 899–913 (2011)
22. Sjøberg, D.I.K., Hannay, J.E., Hansen, O., By Kampenes, V., Karahasanovic, A., Liborg, N.K., C. Rekdal, A.: A survey of controlled experiments in software engineering. IEEE Trans. Softw. Eng. 31(9), 733–753 (2005)
23. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B.: Experimentation in Software Engineering. Springer (2012)
24. Wohlin, C., Runeson, P., da Mota Silveira Neto, P.A., Engstr ́om, E., do Carmo Machado, I., de Almeida, E.S.: On the reliability of mapping studies in software engineering. Journal of Systems and Software 86(10), 2594 – 2610 (2013)
25. Zannier, C., Melnik, G., Maurer, F.: On the success of empirical studies in the international conference on software engineering. In: Proc. of the 28th ICSE. pp. 341–350 (2006)
26. Zhang, M., Hall, T., Baddoo, N.: Code bad smells: A review of current knowledge. J. Softw. Maint. Evol. 23(3), 179–202 (2011)