Using the SCAS Strategy to Perform the Initial Selection of Studies in Systematic Reviews: An Experimental Study

Fábio Octaviano
Federal Institute of São Paulo
Rod. Washington Luís, km 235, CEP 13565-905, São Carlos, SP, Brazil
+55 16 99991 4546
foctaviano@ifsp.edu.br

Cleiton Silva
Federal University of São Carlos
Rod. Washington Luís, km 235, CEP 13565-905, São Carlos, SP, Brazil
+55 16 3351 9488
cleiton.silva@dc.ufscar.br

Sandra Fabbri
Federal University of São Carlos
Rod. Washington Luís, km 235, CEP 13565-905, São Carlos, SP, Brazil
+55 16 3351 9488
sfabbri@dc.ufscar.br

ABSTRACT
Context: Systematic Review (SR) is a well-defined and rigorous methodology used to find relevant evidence about a specific topic of interest. Depending on the number of identified primary studies, the selection activity can be very time-consuming and a strategy, like SCAS, to semi-automate this activity can be helpful. *Objective*: To present an experimental study carried out to evaluate the SCAS strategy. *Method*: We conducted an experiment to compare the efficiency and effectiveness between participants using SCAS and the manual approach. They received necessary training for applying SCAS using tool support. They were divided into five groups, which conducted SRs based on their research areas. *Results*: When applying SCAS, the average effort reduction was 22.33%, and the average percentage error was 3.95% with a minimal loss of 1.6 evidence per SR. In addition, results showed an overall precision of 65.49% on an overall recall of 90.24% when using SCAS. The overall Kappa showed that there is a substantial agreement level between the groups and SCAS. *Conclusion*: The experiment increased the confidence in the strategy, reinforcing that it can reduce the effort required to select primary studies without adversely affecting the overall results of SRs.

Categories and Subject Descriptors
D.2.m [**Software Engineering**]: Miscellaneous – *Evidence-Based Software Engineering (EBSE)*

General Terms
Experimentation, Management.

Keywords
Primary study selection activity; systematic literature review (SLR); evidence-based software engineering (EBSE); Score Citation Automatic Selection (SCAS); experiment; StArt tool.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
EASE '16, June 01-03, 2016, Limerick, Ireland
© 2016 ACM. ISBN 978-1-4503-3691-8/16/06$15.00
DOI: http://dx.doi.org/10.1145/2915970.2916000

1. INTRODUCTION
Software engineering (SE) researchers have increasingly adopted evidence-based software engineering (EBSE) since its introduction in 2004, and systematic reviews (SR) (a.k.a. the systematic literature review (SLR)) have been used to support it, providing mechanisms to identify and aggregate research evidence [1, 2].

Kitchenham and Charters [1] outline three phases to perform a SR: planning, execution, and reporting. During the first phase, reviewers should identify the need for a review and create the review protocol, which contains important information about the SR. In the second phase, they need to identify relevant research and perform the selection of primary studies, data extraction, and data synthesis. Finally, in the third phase, they should report the SR results to relevant communities.

Reviewers from distinct areas within the SE discipline have been adopting SRs. Some challenges have been reported based on reviewers' experiences when undertaking SRs. One potentially problematic aspect of the SR process is the primary study selection [3], which is usually a two-stage process: first, exclude studies, based on reading the titles and abstracts; and then appraise the remaining studies, based on reading the full text. The selection activity can be very time-consuming, in particular for SRs that have large volumes of primary studies to be processed. In this context, the support of automated tools can be useful to minimize this issue.

Marshal and Brereton [4] report some tools available in the literature to support researchers conducting SRs. One of them - StArt (State of the Art through Systematic Review) [5] presents some features to support the decision-making process associated with the initial selection activity: a score used to classify the primary studies based on their relevance (according to the keywords defined in the protocol), and the citation relationships. These features are the basis of the score citation automatic selection (SCAS) strategy, developed to make the initial selection activity as automated as possible [6]. A case study was carried out to preliminarily evaluate the strategy and the results seemed promising, as presented in [6]. Thus, further evaluations are important to investigate the contribution of SCAS and we carried out an experiment reported in this study.

The remainder of this paper is organized as follows: Section 2 presents the background and related work. Section 3 describes the SCAS strategy and the previous results. Section 4 reports the

experiment carried out to provide an additional evaluation of the strategy. Section 5 discusses results and limitations of this work. Finally, Section 6 presents the conclusions.

2. BACKGROUND AND RELATED WORK

Systematic reviews have been used in SE for various research topics [2]. One of the activities associated with the SR process is the selection of primary studies. The selection activity is usually done by considering a large collection of documents (e.g., 750 primary studies in [7] and 653 primary studies in [8]). It is a laborious activity, and is currently performed with minimal automated support. The need for finding mechanisms to assist and speed up this activity is evident, and an automated tool may be very helpful.

In this context, some tools have been developed to support the conduction of SRs in the SE field. Marshall and Brereton [4] identified and classified the most relevant tools. However, the tools have not been properly evaluated and the identified primary studies presented only some case studies or small experiments to exemplify the use of a tool or its effectiveness. Four tools (StArt - State of the Art through Systematic Review [5], SLuRp, SLR-Tool, and SLRTool) support the whole SR process and they are evaluated in [9]. The authors used feature analysis to check the SR activities that each tool supports. In the analysis, it is possible to verify that StArt is the only one that presents a feature to classify primary studies according to their relevance based on a score.

Regarding the non-automatic support for the initial selection activity, three studies have investigated the use of visual techniques within the context of EBSE [10, 11, 12]. They concluded that visual techniques therefore appear to be useful in supporting the SR process. However, these studies do not include or exclude studies automatically; they only support reviewers when performing the selection activity. The use of content-based visual techniques is presented in [10], whilst the use of visualization techniques based on meta-data analysis, such as citation maps and edge bundles, is presented in [11, 12]. Another approach, based on the decisions made by reviewers (two or more) regarding the primary studies, is presented in [13]. Studies should be individually classified as Relevant, Uncertain, or Irrelevant and then categorized, based on the reviewers' decisions, into: A (Relevant / Relevant), B (Relevant / Uncertain), C (Uncertain / Uncertain), D (Relevant / Irrelevant), E (Irrelevant / Uncertain) and F (Irrelevant / Irrelevant). Studies belonging to A and B are indicated for full text reading. Studies belonging to F should be excluded. Studies belonging to D and E should be discussed and classified as A, C or F. Studies belonging to C must have the adaptive full text reading (introduction, conclusion, and so on up to have a decision) to be included or not. However, all the processing is 100% manually, that is, reviewers must review all papers and categorize them to get recommendations.

Regarding the semi-automatic support for the initial selection activity, a strategy using linked data approach for selection process automation in SRs is proposed in [14]. It is based on the use of DBPedia, which is a web data repository that stores information from Wikipedia using structured data (linked data). The strategy begins with the definition of an initial set of relevant studies (called I0) for the SR, and bag of words are extracted from them. Once the reviewers perform the initial selection of a set of studies, the extracted bag of words is enriched by new relevant unknown terms. After that, the remaining studies are automatically processed and compared to the set of relevant studies and, based on an established threshold, the strategy considers a study relevant or not. However, there is not an available tool supporting the strategy. Another approach is presented in [15], whose authors propose two strategies to rank studies according to their importance considering the terms used in the search string: one based on the Vector Model, and the other one based on the simulation of the Boolean expression used in the search string. They are based on the frequency of search string terms found in the title or abstract of a paper. An algorithm was implemented to support the strategies and a case study was carried out to evaluate them. However, there is also no available tool allowing SR reviewers to use or even evaluate the proposed strategies. The third approach, SCAS strategy [6], proposed by the authors of this paper, uses the score feature already implemented and available in the StArt tool, which calculates a score for each study assigning distinct weights for a search term when it is found in the title, abstract, or keywords of a study. In addition, it also considers the number of citations among studies to support the initial selection of studies (initial screening), which is also calculated by StArt. Both features are combined in order to provide a semi-automatic classification of studies. The strategy and the preliminary results are described in Section 3.

It is important to mention that, in the context of this work, the term initial selection (or initial screening) is used to refer to the activity of selecting papers that are relevant to the context of a research topic based on their titles, abstracts, and keywords.

3. THE SCAS STRATEGY

3.1 Strategy Description

The SCAS strategy [6] is based on two features: i) the score, which supports the analysis of primary studies by their content, i.e., the frequency of occurrence of search-string terms in the title, abstract, and keywords; and ii) the number of citations, which shows how many times a study is cited by other studies belonging to the same SR.

The isolated use of either the score or number of citation may be not sufficient to decide the inclusion or exclusion of the primary studies. A study may have a very low score but a high number of citations, or the opposite. In these cases, considering the score or number of citation singly can lead to a hasty decision on the study. Thus, when combined, these complementary features can be used to semi-automate the selection of the primary studies in a SR through the SCAS strategy, which is composed of two phases. In the first phase, the score and citation features are applied to all studies in isolation. In the second phase, the features are combined to suggest if a study should be automatically included, automatically excluded or manually reviewed. Each phase is better discussed next.

3.1.1 Phase 1 - Applying score and citation features

The first step is to apply the score feature. Each study receives a score to represent its relevance to a SR based on its content. Studies with high scores are potentially relevant and should be probably included in an SR, while studies with low scores are potentially non-relevant and should be probably excluded from a SR. To determine if a score is high or low, a cut-off value has to be defined, and two techniques (the 50% rule and the J48 decision tree) can be used to determine the cut-off value.

For the 50% rule, the primary studies are organized in a descending ordered list by score; i.e., the studies with higher scores are positioned at the top of the list. The cut-off value is defined as the score of the study ranked in the middle of the list. For example, if there are forty studies to be analyzed, the score of the twentieth study is used as the cut-off value. If there are an odd number of studies, the quotient should be truncated to zero decimal places in order to obtain the cut-off value. Studies with a score above the cut-off value (high scores) are candidates to be included in a SR. On the other hand, studies with a score below the cut-off value (low scores) are candidates to be excluded from a SR. It is important to highlight that all studies with the same score as the cut-off value are also candidates to be included. The 50% rule was defined based on observations made in the three SRs used in the initial case study presented in Section 3.2.

For the second technique, we generated a J48 decision tree in the Weka tool (available at http://www.cs.waikato.ac.nz/ml/weka/index.html). The J48 decision tree requires at least two input variables (called attributes) and an output variable (called class). The attributes were the score and the number of citations of primary studies. The score must be normalized, and the highest score is set to 1, while all other values are calculated by dividing the score of each study by the highest score. The number of citation is set to 0 (i.e., study is not cited) or 1 (i.e., study is cited at least once). This binary classification was chosen at this stage of our work, but a deeply investigation will be made to determine if there is a better classification for the citation attribute. Based on the score and citation attributes, the class status indicates whether a study should be included or not. We used the decisions made by the experts who conducted the three SRs used in the initial exploratory case study (see Section 3.2) to train the J48 decision tree. Actually, 66% of data were randomly used to train the decision tree and 34% of data were used to execute the decision tree, getting a correctly classified percentage of almost 80%.

The J48 decision tree performs two levels of "pruning". The first is based on the score attribute, i.e., the citation attribute is not considered. The score suggested as a cut-off value is 17.14% of the value of the highest score obtained. The cut-off value should be rounded to the integer value immediately above in the case where the cut-off value has decimal places. For example, if the suggested cut-off value is a score of 21.25 or 21.93, both should be rounded to a score of 22. The second "pruning" is based on the citation attribute, which determines if a study should be included or rejected. However, we do not consider this "pruning" since we judge relevant to investigate the number of citations of the studies with low scores, which are rejected in the first "pruning". Thus, we try to avoid discarding important evidence that would be missed if this set of studies is automatically rejected. Figure 1 shows the resulting J48 decision tree generated in Weka, where it is possible to check the suggested cut-off threshold.

After applying both cut-off techniques, the SCAS strategy determines that the cut-off value should be the lower score between the two chosen candidates (one from the 50% rule and another from the J48 decision tree). Consequently, a larger number of studies is classified with a high score, being considered as potentially relevant.

3.1.2 Phase 2 - Combining score and number of citation features to classify primary studies

The second step is to combine the score and number of

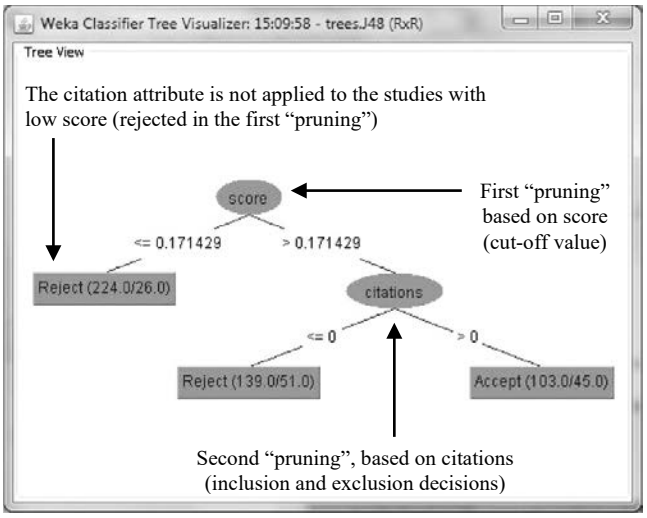

Figure 1. The resulting J48 decision tree in Weka [6].

citations features in order to classify the primary studies into three categories and four quadrants:

- Category 1 ("correct" inclusion – quadrant 1): a study with high score (study positioned above the cut-off value) and that receives at least one citation is probably a relevant study, i.e., a candidate to be included in a SR;
- Category 2 ("correct" exclusion – quadrant 4): a study with a low score (study positioned below the cut-off value) and that receives no citation is probably an irrelevant study, i.e., a candidate to be excluded from a SR;
- Category 3 (studies to be reviewed – quadrants 2 and 3): a study with a high score but no citation, or a study with a low score but at least one citation; these should be reviewed to determine their relevance or irrelevance.

Figure 2 illustrates the classification into quadrants and their corresponding categories (correct inclusion, correct exclusion, and studies to be reviewed) based on the score and number of citations features.

Quadrant 1	**Quadrant 2**
"Correct" Inclusion	**To be Reviewed**
↑ *High Score*	↑ *High Score*
At least 1 Citation	*No Citation*
Quadrant 4	**Quadrant 3**
"Correct" Exclusion	**To be Reviewed**
↓ *Low Score*	↓ *Low Score*
No Citation	At least 1 *Citation*

Figure 2. Combining the score and citation features to define the status of primary studies [6].

In summary, for studies belonging to quadrant 1, the researcher should accept the SCAS recommendation and automatically set these studies to be included based on evidence of a high score and at least one citation. For studies belonging to quadrants 2 and 3, the researcher must manually review these studies in order to set each one to be included or excluded (no automatic action should be taken). For studies belonging to quadrant 4, the researcher can accept the SCAS recommendation and automatically set these

studies to be excluded, based on evidence of a low score and no citation.

3.2 Previous Results

In our previous work [6], we carried out a case study containing three examples in order to demonstrate the use of SCAS strategy. The examples are SRs manually conducted and published in the literature, which vary the topic and the number of primary studies (from dozens to hundreds of studies). These SRs were selected for two main reasons: (i) they were conducted and double-checked by reviewers with experience in conducting SRs; and (ii) they contained all the necessary information (e.g., list of studies included; list of studies excluded; and search-string) to apply our strategy. It is important to highlight that the SRs were not redone since the reviewers provided the original data.

Table 1 summarizes the first example (SR1) and it is organized into two parts: (i) table header – information related to the SR, such as the thematic and number of primary studies; and (ii) primary studies and their score, number of citations, and status (i.e., included or excluded by the "expert" – the researcher(s) who conducted the SR).

For SR1 (presented in Table 1), the "50% rule" suggested that the cut-off value was the score of the paper #48, which was 15. Adopting the "50% rule", all papers with a score of 15 must also be considered. Therefore, two more studies were classified as papers with a high score, which totaled 50 studies. However, the J48 decision tree suggested the cut-off value should be a score of 14, setting the score of the paper #52 as the cut-off value. The worst case of the two techniques, i.e., the result that considers more papers, should be chosen. Consequently, the score of paper #52 was chosen as the cut-off value.

The results of SR1 showed that there were 22 papers classified in quadrant 1 (i.e., high score and receives citation(s)), an indicator that these papers were relevant for inclusion in the SR1. Of this total, 19 were included (papers 2, 3, 5, 6, 8, 11–13, 17, 19, 20, 25, 26, 28, 38, 40, 41, 44, and 48) and three (papers 34, 46, and 51) were excluded by the experts who manually conducted SR1 (see "Author" in Table 1). Papers 52–57, 59–65, 67, 69, 70, 73, 75, 76, and 78–97 (38 in total), had a low score and were not cited – these were classified in quadrant 4, an indicator that these papers should be excluded from SR 1. Only paper 62 was not excluded by the experts. Papers 1, 7, and 10 (quadrant 2 – high score and low citation) and papers 71, 72, and 77 (quadrant 3 – low score and high citation) are some examples of papers that had to be manually reviewed.

Similarly, the same analysis were done for the other two SRs. Table 2 compares the two classifications of the primary studies that comprised the examples used in our exploratory case study: (i) the classification of each primary study into one of the quadrants by the SCAS strategy; and (ii) the classification of each primary study (i.e., included or excluded) by the experts. We assumed that the decisions made by the experts are correct once they are experienced reviewers and conducted or supervised the three SRs published in the literature. Analyzing the data presented in Table 2, it is possible to see that, for SR1, the manual effort required in the first stage of the selection activity, reading the title and abstracts, was reduced by 61.85%; 60 studies of a total of 97 (22 studies belonging to quadrant 1 and 38 belonging to quadrant 4) could be automatically classified using the recommendations of the SCAS strategy. The percentage error was 4.12%; four studies of a total of 97 (three belonging to quadrant 1, and one belonging to quadrant 4) received different classification by the experts who conducted SR1.

Table 1. Classification of studies belonging to SR1 into SCAS quadrants and comparison to the experts' decisions [6]

Study ID	SR 1 Information					
	Title	Authors	Reference	Thematic	Total Included	Total Excluded
	Experimenting with a multi-iteration systematic review in software engineering	Fabiano Cutigi Ferrari José Carlos Maldonado	ESELAW, 2008	Aspect oriented software testing	34	63

Paper ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Score	78	69	65	56	49	46	44	43	38	38	37	35	35	35	32	32	32	32	31	31	31	29	29	28	26	26	25	25	24	23	22	22	22	22	22
Citations	0	1	1	0	35	6	0	18	0	0	9	3	1	0	0	0	7	0	1	1	0	0	0	0	4	9	0	8	0	0	0	0	0	1	0
Expert Decicion	■	□	□	□	□	□	■	□	□	■	□	□	□	■	□	□	□	■	□	□	□	■	□	□	□	□	■	□	■	■	■	■	■	■	■

Paper ID	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70
Score	20	20	20	18	18	17	17	16	16	16	16	16	15	15	15	14	14	13	13	13	12	12	12	11	11	10	10	9	9	9	9	9	9	9	9
Citations	0	0	2	0	10	6	0	0	3	0	2	0	1	0	0	6	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	3	0	0
Expert Decicion	■	■	□	■	□	□	■	□	□	■	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	□	■	■	■	■	■

Paper ID	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97
Score	9	9	8	8	8	8	8	8	8	8	8	6	6	6	6	6	5	5	5	5	5	3	3	3	3	1	1
Citations	7	6	0	3	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Expert Decicion	□	□	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Legend: ▯ Included ■ Excluded □ Quadrant 1 ▒ Quadrant 2 ▒ Quadrant 3 □ Quadrant 4 ○ Cut-off paper

Table 2. Comparison of SCAS and experts classification of studies for the three SRs [6]

SR ID	Total of primary studies	Quadrant 1 "Correct" Inclusion ↑ *High Score* At least 1 *Citation*		Quadrant 2 To be Reviewed ↑ *High Score* No *Citation*		Quadrant 3 To be Reviewed ↓ *Low Score* At least 1 *Citation*		Quadrant 4 "Correct" Exclusion ↓ Low Score No Citation	
SR1	97	**# Total**	22	**# Total**	30	**# Total**	7	**# Total**	38
		# Included*	19	**# Included***	10	**# Included***	4	**# Included***	1
		# Excluded*	3	**# Excluded***	20	**# Excluded***	3	**# Excluded***	37
SR2	37	**# Total**	5	**# Total**	14	**# Total**	3	**# Total**	15
		# Included*	5	**# Included***	9	**# Included***	2	**# Included***	7
		# Excluded*	0	**# Excluded***	5	**# Excluded***	1	**# Excluded***	8
SR3	264	**# Total**	69	**# Total**	63	**# Total**	46	**# Total**	86
		# Included*	33	**# Included***	29	**# Included***	6	**# Included***	6
		# Excluded*	36	**# Excluded***	34	**# Excluded***	40	**# Excluded***	80

The three studies belonging to quadrant 1 required additional reading effort, i.e., three "irrelevant" papers that need to be read by reviewers following recommendations of the SCAS strategy. The paper belonging to quadrant 4 is a false-negative decision, i.e., it is a relevant study that was excluded by the SCAS strategy.

Similarly, the results of SR2 indicated that the effort reduction in the first stage of the selection activity was 54.05%, and the percentage error was 18.91%; 20 studies classified in quadrants 1 and 4 could be automatically decided. Notice that in the case of SR2, which is a tertiary study, cross-citation will likely be low because the gathered studies will not be related to each other in most cases. Finally, the results of SR3 indicated that the effort reduction in the first stage of the selection activity was 58.71% and the percentage error was 15.90%; 155 studies belonging to quadrants 1 and 4 could be automatically classified.

Additionally, in order to measure the agreement between SCAS and the experts, the Cohen's kappa coefficient (a.k.a. Kappa) [16] was calculated for each SR used in the case study. The overall Kappa, calculated based on the data from the three SRs, shows a substantial agreement level, which is a good result according to the interpretation scale in [17].

As the initial results are promising and reveal the usefulness of applying SCAS strategy to support the conduction of SRs, we planned and carried out an experiment to obtain additional results and perform a new evaluation of the SCAS strategy, as presented in the next section.

4. EXPERIMENT DESIGN

4.1 Planning

The main objectives of the experiment are: (i) to evaluate if SCAS strategy is more efficient than the manual approach; (ii) to evaluate the agreement level between SCAS and the reviewers who performed the initial selection; and (iii) to evaluate the effectiveness of SCAS when recommending decisions for studies with decision conflicts.

We planned the experiment using the Goal-Question-Metric (GQM) model [18], as presented in Figure 3. Regarding the metrics, which are the dependent variables of the experiment, it is important to clarify what exactly they intend to measure:

- M1: Time spent for generating the quadrants for the studies using the tool, i.e., the time spent for applying SCAS;
- M2: Time spent by students for manually reviewing the studies belonging to quadrants 1 and 4. We only consider these quadrants because they are the ones that SCAS recommends automatic decisions;
- M3: Number of included studies belonging to quadrant 1 in accordance with group members, i.e., how many studies the reviewers included once SCAS recommendation is for including such studies;
- M4: Number of excluded studies belonging to quadrant 4 in accordance with group members, i.e., how many studies the reviewers excluded once SCAS recommendation is for excluding such studies;
- M5/M6: Number of conflicts (divergences among reviewers on the decision of including or not studies) occurred in the initial selection activity for studies belonging to quadrants 1 and 4, respectively;
- M7: Number of studies with decision conflicts belonging to quadrant 1 that were included by the reviewers after discussion, i.e., how many conflicts were resolved as included once SCAS recommendation is for including such studies;
- M8: Number of studies with decision conflicts belonging to quadrant 4 that were excluded by the reviewers after discussion, i.e., how many conflicts were resolved as excluded once SCAS recommendation is for excluding such studies.

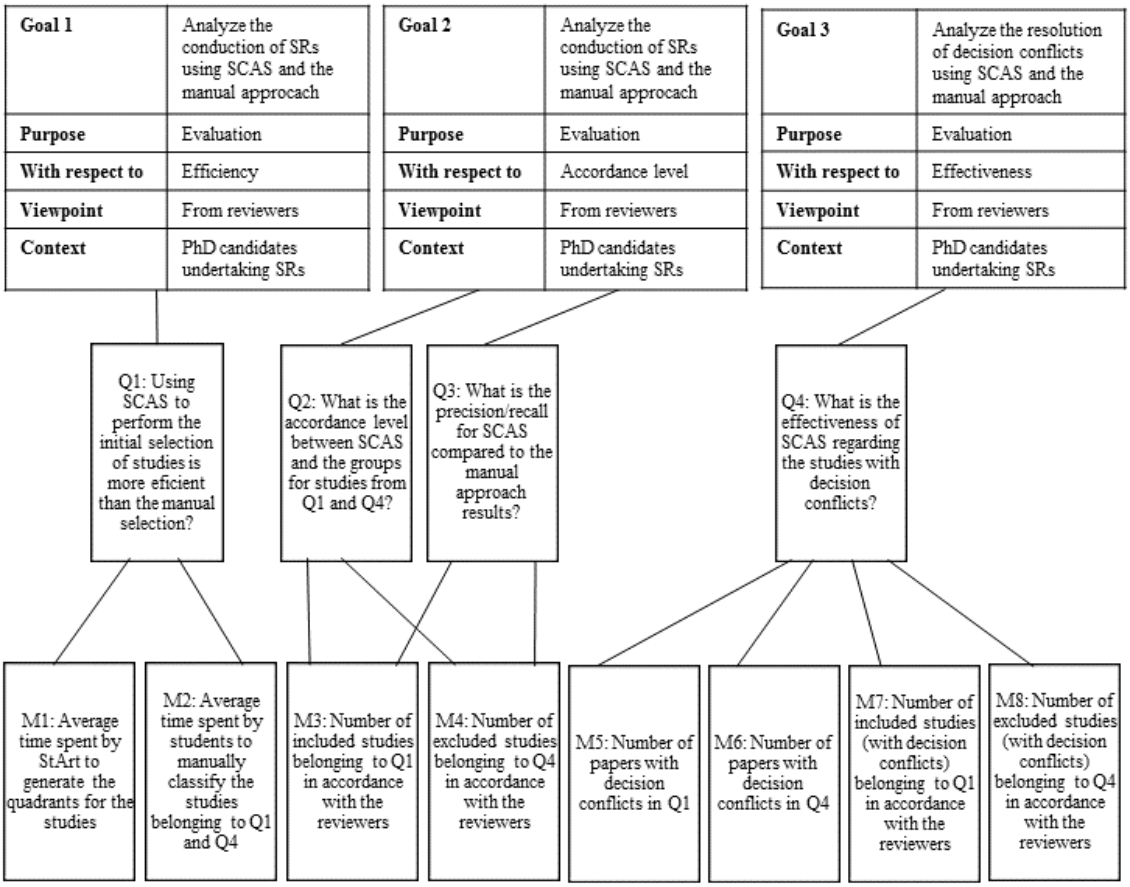

Figure 3. The planned GQM model for the experiment.

4.2 Hypothesis

The null hypotheses of the experiment are:

- H0,1: to perform the initial selection activity using SCAS is not more efficient than a manual initial selection activity;
- H0,2: to perform the initial selection activity using SCAS does not provide similar results as a manual initial selection activity;
- H0,3: SCAS does not help to resolve conflicts among reviewers regarding the decision of including or not studies.

4.3 Population

The population was composed of 21 graduate students (all PhD candidates), during a SR course. Some PhD candidates already knew the SR process and many PhD candidates were introduced to it during the course. There were students from some distinct research areas: 12 from computing (software engineering), five from production engineering, and four from education. They were divided into groups according to their research areas in order to perform SRs for topics of their interest. The software engineering students were divided into three distinct groups according to the specific research subarea they belong to.

It is important to mention that the SR course was multi-disciplinary, offered to some university departments and this is the main reason for having participants not related to the software engineering discipline only. Besides, it would be good to analyze whether the SCAS strategy would provide satisfactory results for other disciplines than software engineering.

4.4 Operation

During the SR course, before the experiment, all PhD candidates were trained on how to perform a SR process and on how to use the StArt tool to support the whole SR process. The course was 64 hours, and 24 of them were used for the SR process and the tool learning. Even the students that already knew the SR process participated in the same training stage. After that, the participants were divided into five groups based on their disciplines of interest, as presented in Table 3.

Table 3. Groups' definition for the experiment

Group id	Members	Research area
1	5	Production engineering
2	4	Education
3	5	Software Engineering
4	4	Software Engineering
5	3	Software Engineering

We carried out the experiment in two days:

- First day:

a) The goals of the experiment were introduced to the participants;

b) Each group chose a distinct topic of interest related to its research area;

c) Each group created a SR protocol and retrieved the primary studies from the online databases they had defined;

d) Each group applied the SCAS strategy for the retrieved primary studies and got the quadrants for the studies, taking note of the time spent for applying SCAS;

e) Each group member manually performed the initial selection activity for the studies belonging to quadrants 1 and 4, reading the titles and abstracts, and making a decision of including or not those studies, taking note of the time spent to perform it. It is important to highlight that the participants were asked to carefully analyze such studies without taking into account the previous classification provided by SCAS;

f) Each group member analyzed and compared the SCAS recommendations for quadrants 1 and 4 to the decision he/she made for the studies belonging to these quadrants;

g) Each group member answered a questionnaire informing its name, research area and group id; the time spent to apply SCAS and to perform the manual selection of the studies belonging to quadrants 1 and 4; the number of included and of excluded studies belonging to quadrant 1; the number of included and of excluded studies belonging to quadrant 4;

- Second day:

a) The group members met and identified the existing conflicts regarding the decisions made for studies belonging to quadrants 1 and 4;

b) After discussion, the group members made a final decision about the studies with decision conflicts;

c) Each group analyzed and compared SCAS recommendations for quadrants 1 and 4 to the final decisions made by the group for the studies belonging to these quadrants;

- A member of each group answered a questionnaire informing the research area and group id; the number of included and of excluded studies belonging to quadrant 1 after the consensus meeting; the number of included and of excluded studies belonging to quadrant 4 after the consensus meeting; the number of studies with decision conflicts belonging to quadrant 1, and how many of them were included and excluded; the number of studies with decision conflicts belonging to quadrant 4, and how many of them were included and excluded;

5. RESULTS AND ANALYSIS

The last task of each group during the experiment was to answer a questionnaire regarding the obtained results. Table 4 presents the answers given by each group after performing the initial selection activity and the consensus meeting. It shows the decisions made for studies belonging to quadrants 1 (Q1) and 4 (Q4), that is, how many primary studies were included and excluded in quadrant 1 and quadrant 4, respectively. We do not consider the studies belonging to quadrants 2 and 3 because they should be manually reviewed using SCAS or not.

We considered the decisions made by SR reviewers as the baseline to compare and evaluate SCAS recommendations. They chose the research topics that they were familiar and had experience on them. Besides, all the final decisions regarding the studies were made after a consensus meeting, where the studies with conflict were discussed to make a final decision.

Regarding the efficiency, answering H0,1, when comparing the time spent for adopting SCAS recommendations with the time spent to perform the manual revision of studies belonging to quadrants 1 and 4, it is possible to see (Table 4) that SCAS is much faster. It is important to mention that the time spent by SCAS may vary depending on the computer configuration where the StArt tool was installed. It is not only related to the number of processed studies. On average, SCAS took around four minutes to run while reviewers took around 95 minutes to complete the initial selection reading titles and abstracts of the primary studies. This means that SCAS is more efficient than the manual revision, as expected.

Considering the number of studies to be assessed in the initial selection, group 1 had an effort reduction of 27.34% (70 studies automatically classified of a total of 256 studies retrieved from databases). Similarly, group 2 had an effort reduction of 22.7%, group 3 had 21.56%, group 4 had 23.41%, and group 5 had 13%.

Table 4. Decisions made for studies belonging to quadrants 1 and 4

	Group ID	1	2	3	4	5
	Research area	Production engineering	Education	Software Engineering	Software Engineering	Software Engineering
	# of studies retrieved	256	260	269	252	154
	Average time spent by SCAS (in min.)	4	6	4	3	2
	Average time spent by reviewers (in min.)	82	105	85	115	88
Q1	**# of studies**	13	15	35	40	10
	# of included studies	9	10	23	24	8
	# of excluded studies	4	5	12	16	2
Q4	**# of studies**	57	44	23	19	10
	# of included studies	2	3	2	0	1
	# of excluded studies	55	40	21	19	9

The percentage error for group 1 was 2.34%; six studies of a total of 256 (four belonging to quadrant 1 and two belonging to quadrant 4) received different classification by reviewers from group 1. The four studies belonging to quadrant 1 are false-positives decisions and would require additional reading effort, i.e., four "irrelevant" studies that need to be read by the reviewers following recommendation of SCAS strategy. The two studies belonging to quadrant 4 are false-negative decisions, i.e., they are relevant studies that were excluded by SCAS strategy. Similarly, the percentage error for group 2 was 3.08% (five false-positives and three false-negatives), for group 3 was 5.2% (12 false-positives and two false-negatives), for group 4 was 6.35% (16 false-positives and no false-negative), and for group 5 was 1.95% (two false-positives and one false-negative). Therefore, the loss of evidence (false-negative decisions) is very low compared to the time saved by using SCAS. Some irrelevant studies were included for full-text reading, but they would be probably excluded just after reading the introduction or conclusion of the full-text files.

Table 5 synthesizes the analysis performed for the data presented in Table 4, in which saved time refers to the time spent by reviewers minus the time spent by SCAS; effort reduction refers to the number of studies that would not be read in the SR; percentage error refers to the number of studies incorrectly classified by SCAS according to the reviewers' decisions; and lost evidence is the number of false-negative decisions.

Table 5. Summary of analysis performed for the groups

Group Id	Saved time	Effort reduction	Percentage error	Lost evidence
1	78 min.	27.34%	2.34%	2
2	99 min.	22.70%	3.08%	3
3	81 min.	21.56%	5.20%	2
4	112 min.	23.41%	6.35%	0
5	86 min.	13.00%	1.95%	1

Additionally, in order to measure the agreement between SCAS and the groups, the Cohen's kappa coefficient [16] was calculated. Kappa is calculated by the equation $k = (Pr(a) – Pr(e)) / (1-Pr(e))$, where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category.

If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr(e)), $\kappa < 0$. Table 6 presents the interpretation for Kappa values suggested in [17].

Table 6. Interpretation for Kappa values [17]

Values of Kappa	Interpretation
<0	No agreement
0.00-0.19	Poor agreement
0.20-0.39	Fair agreement
0.40-0.59	Moderate agreement
0.60-0.79	Substantial agreement
0.80-1.00	Almost perfect agreement

Aiming to apply Kappa for comparing the agreement level between SCAS and the groups, we made the following assumptions: i) raters - the group's decisions and the SCAS recommendations; ii) categories - included and excluded; and iii) observed data - the primary studies considered belonging to quadrant 1 and quadrant 4, as they contain the studies that SCAS recommends to include or to exclude automatically. Quadrants 2 and 3 were not considered as researchers should manually review the studies belonging to them, as explained before.

Table 7 presents the calculated values for each group (manual review and SCAS), and their interpretations according to Table 6. Based on the data shown in Table 7, it is possible to verify that groups 1, 2, and 5 achieved a substantial agreement level compared with SCAS recommendations. However, the same interpretation was not true for groups 3 and 4 that achieved a moderate agreement level. The overall Kappa, calculated based on data from the five groups, shows a substantial agreement level, which is a good result.

For further analysis of SCAS, we calculated the precision and recall for each SR performed by the groups. In information retrieval with binary classification, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. In our experiment, as the goal is to evaluate quadrants 1 and 4, for which SCAS recommends automatic decisions, precision and recall were calculated for the SRs considering only the studies belonging to these quadrants, as presented in Table 8, which also includes the overall precision and recall considering data from the five groups.

Table 7. Values and interpretation of Kappa for the groups

Group	Kappa	Interpretation
1	0.70	Substantial agreement
2	0.62	Substantial agreement
3	0.53	Moderate agreement
4	0.49	Moderate agreement
5	0.70	Substantial agreement
Overall	0.62	Substantial agreement

Table 8. Precision/recall for the SRs undertaken by the groups

Group id	Precision	Recall
1	69.23%	81.82%
2	66.67%	76.92%
3	65.71%	92%
4	60%	100%
5	80%	88.89%
Overall	65.49%	90.24%

Thus, answering H0,2, based on the low average percentage error (see Table 5), on a calculated Kappa saying that the overall agreement level between SCAS and reviewers is substantial (see Table 7), and on the good overall recall (see Table 8), we can deduce that SCAS provides similar results than a totally manual initial selection activity.

The decision conflicts that happened for quadrants 1 and 4 were also analyzed in order to verify whether SCAS would be helpful in case of divergence of opinions among reviewers. Table 9 shows the decision conflicts reported by each group and the corresponding decisions they made after discussing. Considering only the conflicts, the correctness rate (SCAS recommendations and reviewers' decisions are the same) for group 1 was 80%, that is, 16 out of 20 conflicts (three for quadrant 1 and 13 for quadrant 4) were resolved according to SCAS recommendations. The correctness rate for group 2 was 50%, for group 3 was 60%, for group 4 was 53.85%, and for group 5 was 25%. The results, except for group 1, are not good enough to prove that SCAS is helpful in resolving decision conflicts. However, when we analyze Tables 4 and 9 together, it is possible to note that all false-negative, that are the worst case because it implies in loss of evidence, refer to studies in conflict among group members. This means that SCAS would miss some studies that had no total agreement among reviewers. Thus, we consider, at this stage of our work, that these studies probably are not the most relevant studies of the SRs because of the reviewers' disagreement and because they got low scores once they were classified into quadrant 4.

Thus, answering H0,3, based on data from Table 9, SCAS was correct in 58.89% of suggestions made for the conflict resolutions, which is not enough to prove that SCAS is very helpful to resolve decision conflicts.

Finally, considering that completeness is critical for SRs, we would like to highlight that even humans can make mistakes and exclude relevant studies in the selection activity. Bad titles or badly written abstracts can be a cause of such mistakes.

6. VALIDITY EVALUATION

Based on the threats to validity mentioned in [19], we highlight:

- Internal and construction validities: the research topics and the participants' experience level are threats, once they possibly started their roles as researchers in different periods, come from distinct areas, and also had distinct experience level in performing SRs. We tried to minimize these threats by selecting graduate students (PhD candidates) who are performing researches, considering they are in an acceptable maturity level as researchers. In addition, they participated in a course of 64 hours about the whole SR process for standardizing the knowledge level in performing SRs, and chose research topics they were familiar and able to assess studies.
- Conclusion validity: the conduction of a SR may be a subjective process, being influenced by the participants' profiles and experience level, and by the level of understanding that they acquired from the training stage. In addition, the participants performed the manual classification of studies having a previous knowledge of the classification made by SCAS. Trying to minimize this threat, the decisions made by the group members were compared among them in a consensus meeting. We compared the percentage error between the SCAS recommendations and the decisions made by the groups, and evaluated the agreement level through the Kappa coefficient, but only after the groups having the final decisions regarding the studies.
- External validity: it is plausible to say that the obtained results could be different in another set of participants. Trying to minimize this threat, we chose a population of researchers composed of PhD candidates, who are supposed to have an acceptable level of maturity in conducting researches. Besides, the generalization of our results is subject to certain limitations, mainly because only three topics of software engineering were analyzed, as well as just two research areas but software engineering (production engineering and education).

7. SUMMARY AND FUTURE WORK

Reviewers often face a large number of primary studies for performing the initial selection activity in SRs. In this context, SCAS strategy help in minimizing the effort required to complete it by suggesting studies to be automatically included or excluded. The main contribution of this paper is to present an experimental study for evaluating the strategy, in addition to the previous results reported in [6] and reported in Section 3.2.

The results of the experiment showed that, on average, the effort reduction was 22.33% when automatically accepting studies classified in quadrants 1 and 4, and that the percentage error was 3.95%, with a short loss of evidence, which is equal to 1.6 studies (false-negatives) per SR. This means that the effort reduction is significant compared to the percentage error when accepting the quadrants 1 and 4 automatically.

Table 9. Conflicts detected in quadrants 1 and 4 and their resolutions

	Group ID	1	2	3	4	5
Q1	**# of studies**	13	15	35	40	10
	# of conflicts	5	7	18	23	3
	# of conflicts resolved as inclusion	3	0	12	11	1
	# of conflicts resolved as exclusion	2	7	6	12	2
Q4	**# of studies**	57	44	23	19	10
	# of conflicts	15	13	2	3	1
	# of conflicts resolved as inclusion	2	3	2	0	1
	# of conflicts resolved as exclusion	13	10	0	3	0

Besides, all lost evidence referred to studies with decision conflicts among reviewers, and were not considered relevant by all group members. In addition, we performed the precision and recall analysis considering SCAS recommendations for quadrants 1 and 4 for the five SRs undertaken by the groups in the experiment. Overall results showed a precision of 65.49% on a recall of 90.24%. The overall Kappa was calculated and showed that there is a substantial agreement level between the reviewers and SCAS. Thus, it is possible to say that the presented results reinforce the usefulness of using SCAS to support SRs.

We also evaluated if SCAS would be useful considering only the conflicting studies. Results showed that SCAS was correct in 58.89% of suggestions made for the conflict resolution, which is not enough to prove its effectiveness in this specific case. However, SCAS could be used as the role of an additional reviewer once it recommends the inclusion or exclusion of studies, in case of a reviewer does not feel comfortable to automatically apply SCAS recommendations.

The limitations of this work are better explained in Section 6, where the threats to validity are presented and the actions taken by the authors to minimize them.

As future work, we intend to: (i) investigate how to use the time of publication dimension in addition to the number of citations of studies, trying to determine a citation coefficient for improving the strategy; (ii) keep evaluating SCAS through new experiments or case studies aiming, among other points, to evaluate the required effort for processing false-positives, and to deeply investigate the main causes of false-negatives; and (iii) get feedback on the strategy from StArt users once they apply SCAS in their SRs, which is possible through the StArt Online Community available at the official StArt webpage: http://lapes.dc.ufscar.br/tools/start_tool.

8. REFERENCES

[1] Kitchenham, B. and Charters, S. 2007. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University.

[2] Kitchenham, B. and Brereton, P. 2013. A systematic review of systematic review process research in software engineering. *Information and Software Technology* 55 (Aug. 2013), 2049-2075.

[3] Zhang, H. and Babar, M. A. 2011. An empirical investigation of systematic reviews in software engineering. In *Proceedings of the Int. Symposium on Empirical Software Engineering and Measurement* (Banff, Canada). IEEE, 1–10.

[4] Marshall, C. and Brereton, P. 2013. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In *Proceedings of the International Symposium on Empirical Software Engineering and Measurement* (Baltimore, United States). IEEE, 296-299.

[5] Fabbri, S., Hernandes, E., Di Thommazo, A., Belgamo, A., Zamboni, A., and Silva, C. 2012. Using information visualization and text mining to facilitate the conduction of systematic literature reviews. In *Proceedings of the 14th International Conference on Enterprise Information* (Wroclaw, Poland). Springer, 243-256.

[6] Octaviano, F., Felizardo, K., Maldonado, J., and Fabbri, S. 2015. Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable? *J. Empirical Software Engineering* 20, 6. (Dec. 2015), 1898-1917. DOI= 10.1007/s10664-014-9342-8.

[7] Shaw, M. and Clements, P. 2006. *The golden age of software architecture: A comprehensive survey*. Technical Report CMU-ISRI-06-101. Software Engineering Institute, Carnegie Mellon University.

[8] Shepperd, M. 2007. Software project economics: A roadmap. In *Proceedings of the Workshop on the Future of Software Engineering* (Minneapolis, United States), IEEE, 304–315.

[9] Marshall, C., Brereton, P., and Kitchenham, B. 2014. Tools to Support Systematic Reviews in Software Engineering: A Feature Analysis. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (Torino, Italy), 139-148.

[10] Malheiros, V., Hohn, E., Pinho, R., Mendonca, M., and Maldonado, J. 2007. A visual text mining approach for systematic reviews, in *Proceedings of the International Symposium on Empirical Software Engineering and Measurement* (Madrid, Spain). IEEE, 245-254.

[11] Felizardo, K., Salleh, N., Martins, R., Mendes, E., MacDonell, S., and Maldonado, J. 2011. Using visual text mining to support the study selection activity in systematic literature reviews. In *Proceedings of the International Symposium on Empirical Software Engineering and Measurement* (Banff, Canada). IEEE, 1-10.

[12] Felizardo, K., Andery, G., Paulovich, F., Minghim, R., and Maldonado, J. 2012. A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology* 54, 10 (2012), 1079-1091.

[13] Ali, N. B. and Petersen, K. 2014. Evaluating strategies for study selection in systematic literature studies. In *Proceedings of the International Symposium on Empirical Software Engineering and Measurement* (Torino, Italy), ACM. DOI=10.1145/2652524.2652557.

[14] Tomassetti, F., Rizzo, G., Vetro, A., Ardito, L., Torchiano, M., and Morisio, M. 2011. Linked data approach for selection process automation in systematic reviews. In *Proceedings of the Conference on Evaluation Assessment in Software Engineering* (Durham, United Kingdom), 31-35.

[15] Abilio, R., Vale, G., Pereira, D., Oliveira, C., Morais, F., and Costa, H. 2014. Systematic literature review supported by information retrieval techniques: A case study. In *Proceedings of the Latin American Computing Conference*, (Montevideo, Uruguay), 1-11.

[16] Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22, 2 (1996), 249–254.

[17] Landis, J. and Koch, G. 1977. The measurement of observer agreement for categorical data. *J. Biometrics* 33, 1 (1977), 159-174.

[18] Basili, V., Caldiera, G., and Rombach, H. 1994. Goal Question Metric Approach. *Encyclopedia of software Engineering*, 527 – 532.

[19] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., and Wesslén, A. 2000. *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, Boston, USA.