A Map of Threats to Validity of Systematic Literature Reviews in Software Engineering

Xin Zhou
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
job@wetist.com

Yuqin Jin
College of Information Technology
Nanjing University of Chinese Medicine
Nanjing,Jiangsu,P.R.China
yqjin@njucm.edu.cn

He Zhang
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
hezhang@nju.edu.cn

Shanshan Li
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
dreamhigh_ss@sina.com

Xin Huang
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
njuhuangx@outlook.com

***Abstract*—Context: The assessment of Threats to Validity (TTVs) is critical to secure the quality of empirical studies in Software Engineering (SE). In the recent decade, Systematic Literature Review (SLR) was becoming an increasingly important empirical research method in SE. One of the mechanisms of insuring the level of scientific value in the findings of an SLR is to rigorously assess its validity. Hence, it is necessary to realize the status quo and issues of TTVs of SLRs in SE. Objective: This study aims to investigate the-state-of-the-practice of TTVs of the SLRs published in SE, and further support SE researchers to improve the assessment and strategies against TTVs in order to increase the quality of SLRs in SE. Method: We conducted a tertiary study by reviewing the SLRs in SE that report the assessment of TTVs. Results: We identified 316 SLRs published from 2004 to the first half of 2015, in which TTVs are discussed. The issues associated to TTVs were also summarized and categorized. Conclusion: The common TTVs related to SLR research, such as internal validity and reliability, were thoroughly discussed in most SLRs. The threats to construct validity and external validity drew less attention. Moreover, there are few strategies and tactics being reported to cope with the various TTVs.**

Keywords*—*Systematic (Literature) Review, Threats to Validity, Evidence-Based Software Engineering

I. INTRODUCTION

Kitchenham, Dyba and Jorgensen et al. introduced Evidence-Based Software Engineering (EBSE) in 2004 [1]. Systematic Literature Review (SLR) has become an important research methodology in Empirical Software Engineering (EMSE) since 2004. One critical element in applying this methodology is to design and mitigate Threats to Validity (TTVs). Researchers can draw on a wide range of resources to support SLR, in particular, considering TTVs: generic TTV lists in literatures [2][3] and specific TTVs reported in similar empirical report.

Validity is a property of inference and SLRs involve a range of TTVs. The definitions of common TTVs are listed in TABLE I.

There is no existing systematic review of the validity assessment mechanisms for SLRs in SE. Many researchers report their TTVs as one of the components of reporting an SLR without a good understanding of how to effectively perform validity assessment. It can be a significant challenge to perform a rigorous validity assessment of SLRs.

TABLE I. Definition of Validity

Category	Definition
Construct Validity	Identify correct operational measures for the concepts being studied
Internal Validity	Seek to establish a causal relationship, whereby certain conditions are believed to lead to other conditions, as distinguished from spurious relationships.
External Validity	Define the domain to which a study's findings can be generalized.
Conclusion Validity	Demonstrate that the operations of a study such as the data collection procedure can be repeated, with the same results.

The main contribution of this research is that it presents a ten-year overview of the analysis of TTVs in SLRs by SE researchers. We identified the SLRs that provide definite presentation about their TTVs (e.g. the paper includes the section to discuss TTVs or limitation), analyzed the aspects and categories of existing TTVs, their probable influence and control tactics, and compiled a list of the most often used control tactics for each aspect. For each threat and its control tactics, researchers can gain the knowledge of the most concerned criteria, which is particularly helpful in research process and applying validity assessment for future SLRs.

Our research aims to help researchers to realize the state-of-the-practice of TTVs in SE and enable their development on the body of knowledge about rigorous validity assessment systems, which can inevitably improve the scientific value of the findings and conclusions of SLRs.

In order to understand the validity assessment practices used in the reported SLRs, we carried out a review of the existing SLRs that discuss the TTVs or limitations of the papers. Due to the fact that SLRs and Systematic Mapping Studies (SMS) may have similar aspects or control tactics

1530-1362/16 $31.00 © 2016 IEEE
DOI 10.1109/APSEC.2016.62

of the TTVs, this research analyzes the aspects and control tactics of the TTVs in SLR and SMS comprehensively.

All validity assessment systems (e.g. check-lists, guidelines, and abstract) used in SLRs in SE have been enumerated in this paper. Hence, this work offers a valuable reference for the future researchers conducting SLRs. We also identified and classified categories of common TTVs of the existing SLRs and analyzed their probable influence and control tactics found in different TTVs categories.

The rest of this paper is structured as follows. Section 1 describes the introduction and background. Section 2 describes the research methodology of this study. In Section 3, the selected SLRs in this study are discussed. Section 4 elaborates related work. Section 5 discusses the threats to validity. Section 6 discusses the future work and draws the conclusion.

II. Method And Result

Our review adopted the comprehensive guideline specified by Kitchenham and Charters in collecting the search strategies [2].

A. Research Questions

The research questions for our study are shown as follows：

- RQ1: What are the common TTVs identified/reported in the SLRs in SE?
- RQ2: What consequences/influences are related to the identified TTVs currently?
- RQ3: What strategies and tactics were used to mitigate the TTVs in the SLRs in SE?

B. Search Strategy

In this review, the approach of Quasi-Gold Standard (QGS) [4][8] was adopted. Our search consists of three stages: manual search, automatic search and snowballing. The details of each search stage are described below. We searched SLRs between 2004 and first half of 2015.

1) Manual Search

The manual search was initiated in 2015. At this stage, the authors jointly chose the venues (e.g. journals, conference proceedings) recognized as very specific to Empirical Software Engineering (ESE) and Evidence-Based Software Engineering (EBSE) as well as famously generic publication venues in SE. After carefully considering the venues available in SE community, the authors selected six of them for manual search.

a) Conferences:

International Symposium on Empirical Software Engineering and Measurement (ESEM), Evaluation and Assessment in Software Architecture (EASE)

b) Journals:

Empirical Software Engineering (EMSE), IEEE Transactions on Software Engineering (TSE), Information & Software Technology (IST) and Journal of Systems & Software (JSS)

2) Automated Search

The automated search was conducted through four of the major publishers' digital library portals [8]: IEEE Xplore, ACM Digital Library, Science Direct, and Springer Link. We restricted the search scope between 2004 and the first half of 2015. Searching the fields of title, keyword and abstract of the publications performed the search. We reused using the following search string from previous work [5], and checked the result.

(software AND (((systematic OR controlled OR structured OR exhaustive OR comparative OR evidence) AND (review OR survey OR (literature search) OR map)) OR (mapping study) OR (scoping study) OR (systematic map) OR (tertiary study) or meta-analysis))

Sensitivity is an important metric for evaluating the quality and efficiency of a search strategy. The corresponding quasi-sensitivity reached 85.6% and quasi-precision reached 1.48%. It confirms the search string is still valid for carrying out automated search.

3) Snowballing

Although the automated search covers the majority of SE publications, we might still miss some SLRs. To identify as many SLRs as possible, we further employed the snowballing strategy to seek more SLRs. We used Google Scholar to check the papers with reference to the three EBSE seminal papers [1][13][14] and two versions of the guidelines on SLRs in SE [2][12].

The search strategy was determined by three researchers and reviewed by our supervisor to ensure not to miss any study. We assigned the workload to make sure that each study was checked by at least two researchers independently to minimize the potential impact of any bias. The disagreements on search and selection opinions were solved in joint discussion. Any disagreements that could not be solved were escalated to the supervisor.

C. Inclusion and Exclusion

With respect to the objectives and research questions, the following inclusion and exclusion criteria were applied:

- Inclusion criteria:

I1. The abstract or title explicitly states that the article is a type of systematic literature review.
I2. The paper is in the area of software engineering.
I3. The paper is peer-reviewed (journal article, conference paper).
I4. The paper is a regular paper or a full paper.
I5. The paper presents TTVs related issues as a part of SLR report.
I6. The full-text of the paper is available.

- Exclusion criteria:

E1. The paper is not written in English.
E2. After using a variety of solutions, the paper's full-text is not accessible.
E3. The paper is a gray publication without peer review, e.g., technical reports.
E4. The paper is explicitly a short paper or with less than six pages.

All of these criteria were used to identify the SLRs, which contain TTVs. Based on these criteria, we excluded grey literature (i.e., non peer-reviewed). The papers with fewer than six pages were excluded because we observed that those papers could not contain sufficient details about the procedure of performing SLRs, especially TTVs.

D. Data Extraction

The data extracted from each selected SLR are listed in TABLE II.

TABLE II. Data extraction form

Attribute	Description
Title	The title of the selected paper.
Year	The publication year of the selected paper
Study Type	The type of the study:(Meta analysis=1, SLR =2 or Mapping study=3.)
If there is a discussion on the TTVs?	Whether the selected paper mentioned the Threats to Validity? (Yes = 1 / Possible threats are existed, but its effects are not discussed = 0.5 / No = 0);
Country	The countries where authors' affiliation are situated
Category of TTVs	What kind of Threats to Validity the paper presented? (Construct, Internal, External, Conclusion, Content, Concurrent, Predictive, Statistical, Others)
Influence(s) of TTVS	What influence(s) may be associated with the TTVs
Strategy of TTVs	Which strategies were used to mitigate the TTVs.

E. Data Synthesis

The results of our work were further synthesized based on categorization. For this reason, we performed thematic synthesis [21] to answer the RQ1 and RQ2. To answer the RQ3, we need to get a final strategy or tactics to relieve the influence. We collected information about how to solve the influence and then translated the data we extracted from included SLRs to a new interpretation which can be used to relieve the influence generally. The meta-ethnography method [22] was adopted here.

F. Result

Because of the large sample size, we conducted a pilot. During the pilot, we established some relevant identifying information: such as the "threats to validity", "reinforcement and weakness", "limitation", etc. The selection steps in our work are: Firstly, we retrieved 256 SLRs between 2004 and the first half of 2015 in manual search stage; secondly, the results of the automated search stage are listed in TABLE III; thirdly, we identified 314 SLRs in snowballing stage. After inclusion and exclusion stage, we remove duplicate SLRs. Finally, 316 SLRs remain to be extracted and analyzed.

TABLE III. Summary of search results

Database	No. of retrieved SLRs	No. of selected SLRs
IEEE Xplore	1498	232
ACM DL	206	87
Science Direct	578	168
Springer	19465	143
Total	21747	630

Considering the publication year of these SLRs, We observe that an increasing number of SLRs report the details of the TTVs (Figure 1).

Among the selected 316 SLRs, there are 178 systematic literature reviews, 132 systematic mapping studies and 6 Meta-analyses.

Considering their publication venues, the top six venues are IST (72), JSS (29), EASE (25), ESEM (5), TSE (4) and EMSE (3). These six venues are all recognized as highly specific to empirical and evidence-based software engineering or highly reputed generic publication venues in SE. In particular, the journal of Information and Software Technology (IST) dominated in all SE publication venues.

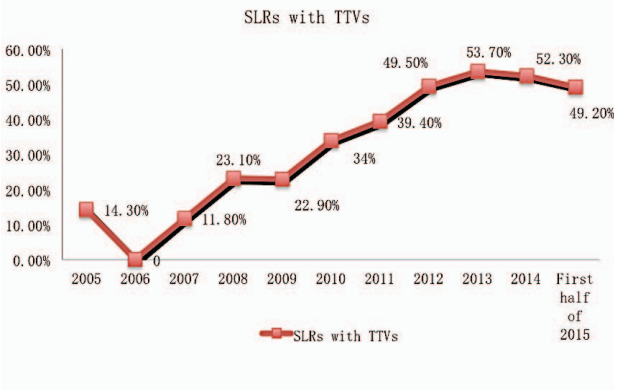

Figure 1. Distribution of Publication Years of SLRs

According to the extracted data of total 316 selected SLR papers, most authors focused on the conducting phase are 87% (275). The other two phases (planning and reporting) are 6% (19) and 7% (22) respectively.

III. DISCUSSION

In this section, we discuss the research questions based on the data from the 316 papers. If you want to obtain more information (e.g., the full list), please contact us by email.

A. RQ1: Common Threats to Validities

The validity assessment is an essential part of any empirical study, including SLRs. The validity commonly includes Construct Validity, Internal Validity, External Validity, Conclusion Validity, etc. Other validities (Theoretical validity, Interpretive validity) were rarely mentioned in the field of software engineering, so they are not discussed in this paper. Based on our work, we classified the identified threats to the four categories of validities shown

in TABLE IV. Note that one threat may be associated with multiple validities.

After data extraction, we classified the aspects of TTVs. SLR researchers discussed these TTVs in their studies (as shown in TABLE IV). Frequency means that the number of each threat appearing in the SLRs. We distinguished these threats based on the research phases of SLR. These threats could be described as follows:

Planning Phase:

Non-specification of SLR's setting and sufficient details: If the SLR's setting (venues, search string, etc.) or other important details are not clearly stated, then this may pose a threat to repeatability and replicability of the study.

Incorrect or incomplete search terms in automatic search: The search string in the searching process may include inadequate search terms related to research topic. Sonia also applied terms patterns and adapted the search string to each digital library with the purpose of making easier the replication of the process [15].

Lack of standard languages and terminologies: It means SLRs used different terms for similar concepts. For example, if the basic concept of coupling is not well understood, then the metric that captures coupling may be inaccurate or incorrect.

Inappropriate search method: Researchers apply search methods to search for relevant papers to get evidence for their studies. But imperfect methods may lead to missing relevant papers (e.g. using automatic search only).

Incomprehensive venues or databases: The library resources used to search for primary studies do not include some important resources databases.

Inappropriate inclusion & exclusion criteria: The researchers are not familiar with this research field, so they may put forward some unsuitable inclusion & exclusion criterion.

Inappropriate research questions: The researchers are not familiar with this research field, so they may put forward some unsuitable research questions.

Inadequate size and number of samples: If the sample size is inadequate or too small to be reasonable, then the validity of the results is not assured.

Restricted time span: Researchers cannot anticipate other relevant studies outside the time span, constructively by limited effort.

Culture bias: Due to the authors of the cultural differences, the validity of the results is not assured. (e.g., preferences for the studies of some researchers' nationality)

Conducting Phase:

Bias in Study Selection: In the process of search, reviewers have own subjective conjecture, and they do not completely use the inclusion and exclusion criteria for judgment.

Identification error of primary studies in the searching process: It means some errors (e.g., related studies are not chosen or irrelevant studies are chosen), which may be found in the search process.

Incomplete research information in primary study: SLRs do not present objective and detail information.

Paper/database inaccessible: Some researchers could not download the papers and contact the author(s) of the exact article/some digital libraries which they do not have the authority to access.

Misclassification of primary study: Researchers may not consider all the classifications at the beginning of the research process or the primary studies selected by the reviewers are mistakenly classified.

Primary study duplication: The duplication of papers is a potential threat (e.g., the same paper included in more than one database or in more than one journal). Researchers need to identify and remove the duplication [16]. A study was reported by more than one paper.

Publication bias: It refers to the problem that positive results are more likely to be published than negative results [2]. Among our selected SLRs, there are some concrete forms of ***publication bias***. Some SLRs may have a tendency to report particular kinds of success factors. Pacheco and Garcia considered the gray literature (technical reports, work in progress, unpublished or non peer-reviewed publications) was a publication bias [17]. However, by checking the guideline [2] and from our recognition, we find this is misunderstanding. Unpublished or not peer-reviewed publications always have the problem that ignore negative results and they should paid more attention to positive results. So we cannot identify the grey literature as publication bias.

Bias in Data Extraction: In the process of data extraction, reviewers do not completely understand the definition of data extraction item and the relationship with research questions.

Subjective interpretation about the extracted data: Researchers may have different interpretation of the extracted data and also different opinions on how to deal with the data.

Unsatisfactory Data Synthesis: Synthesis may be complete and it merely presents our preliminary synthesis.

Subjective quality assessment: It refers to authors' assessment of the criteria based on their own judgment, which may lead to bias and can be a threat. Heckman and Williams stated that their first author assessed the quality of the papers based on 10 questions, which may lead to measurement bias [18].

Reporting Phase:

Primary study generalizability (Concentrate in a narrow area): It is found in only one included SLR [20]. The low primary study generalizability may lead to low generalizability of the SLR conclusion.

Lack of expert evaluation: The conclusions or results should be evaluated by an expert to understand and interpret

their true meaning and significance [19]. Without expert assessment there might be erroneous conclusions reported.

B. RQ2: Consequences and Influences

Obviously, untreated TTVs will reduce the quality of SLRs. One of the reliable mechanisms of ensuring the level of scientific value in the findings of an SLR is to rigorously assess the validity of the SLR. According to the data extracted, there may exist various influences caused by different TTVs (Figure 2).

The most common and important influence is from missing of relevant primary studies' which can result threats like ***Incorrect search method***, ***Incorrect or incomplete search terms in automatic search***, ***Restricted time span***, ***Identification error of primary studies in the searching process***, ***Paper/database inaccessible***, ***Bias in data extraction***, ***Bias in study selection***, ***Publication bias*** and ***Incomprehensible venues or database***. Missing of relevant primary studies will seriously impact the quality of SLRs. About 11% of included SLRs (28) reported the possibility of this influence. Renato et al. indicate that they could not guarantee all related primary studies were selected [S42].

As illustrated in Figure 2, threats may associated with more than one kind of influence. Mentioned threats in RQ1, apart from the influence of missing relevant primary studies, some of them also lead to other influences. ***Incorrect search method*** may cause the exclusion of right papers. ***Incorrect or incomplete search terms in automatic search*** can also limit the variety of information available via search engines. ***Bias in data extraction*** may result in the incorrect classification of publications.

Bias in study selection will have influence on inaccuracy of data and incorrect classification of publications. In addition, the latter influence can be caused by ***Subjective interpretation about the extracted data***.

Non-specification of SLR's setting and sufficient details pose a threat to repeatability and replicability of the study. SLR's setting means that SLR is conducted based on guideline.

Inadequate size and number of samples may bring about the influence that the validity of the results is not assured.

Primary study duplication and ***Lack of expert evaluation*** will have influence on statistics and the accuracy of conclusion. The consequence of ***Unsatisfactory data synthesis*** may lower the quality of the data analysis and ***subjective quality assessment*** will affect the impartiality of the quality evaluation.

Misclassification of primary study may affect the impartiality of the extracted data. ***Inappropriate inclusion & exclusion criteria*** are possible to create different understandings among the reviewers [S34]. ***Inappropriate research questions*** and ***Lack of standard languages and terminologies*** will make chaos in the process of research.

TABLE IV. Common Threats & Mapping of Validity

Validity	Aspect of Threats	Frequency
Ct, In	Non-specification of SLR's setting and sufficient details	7
Ct, In	Inappropriate or incomplete search terms in automatic search	91
Ct, In	Lack of standard languages and terminologies	10
Ct, In	Incorrect search method	39
Ct	Incomprehensible venues or database	68
Ct	Inappropriate inclusion & exclusion criteria	32
Ct	Inappropriate research questions	5
In	Inadequate size and number of samples	3
Ct, Et	Restricted time span	19
In	Culture bias	4
In, Cn	Bias in study selection	122
Ct, In, Cn	Identification error of primary studies	8
Et	Incomplete research information in primary study	28
Et	Paper/database inaccessible	5
In, Cn	Misclassification of primary study	65
In, Cn	Primary study duplication	9
In	Publication bias	49
In, Cn	Bias in data extraction	121
In, Cn	Subjective interpretation about the extracted data	19
In	Unsatisfactory data synthesis	18
In	Subjective quality assessment	38
Et	Primary study generalizability	13
Ct, In	Lack of expert evaluation	3

(Construct validity: Ct, Internal validity: In, External validity: Et, Conclusion validity: Cn)

C. RQ3: Strategies and Tactics

There is no doubt that evaluating the TTVs of an SLR's results is very important. Some strategies are as follows:

Non-specification of SLR's setting and sufficient details: Establish the protocol for the study, meanwhile hold specification of SLR's setting and relevant details.

Incorrect or incomplete search terms in automatic search: 1) All decisions and results need to be checked, rechecked and inconsistencies will be resolved [S13]. 2) Including additional terms in the search string and searching additional libraries [S24]. 3) Developing a tool to extract citation information that considers the peculiarities of each search engine, reducing the number of possible mistakes.

Lack of standard languages and terminologies: We can use the external evaluation.

Incorrect search method: 1) Combine the method of automatic search and manual search [S16]. Balance the recall and precision of the search string and collect publications from different sources [S32]. 2) The irrelevant papers are excluded after reading the title, abstract and conclusion [S38].

Incomprehensive venues or database: 1) Use a larger database that includes inaccessible database [S1]. 2) Use multiple databases and tools that executed the queries on the different sources to reduce subjective errors during the search phase [S8].

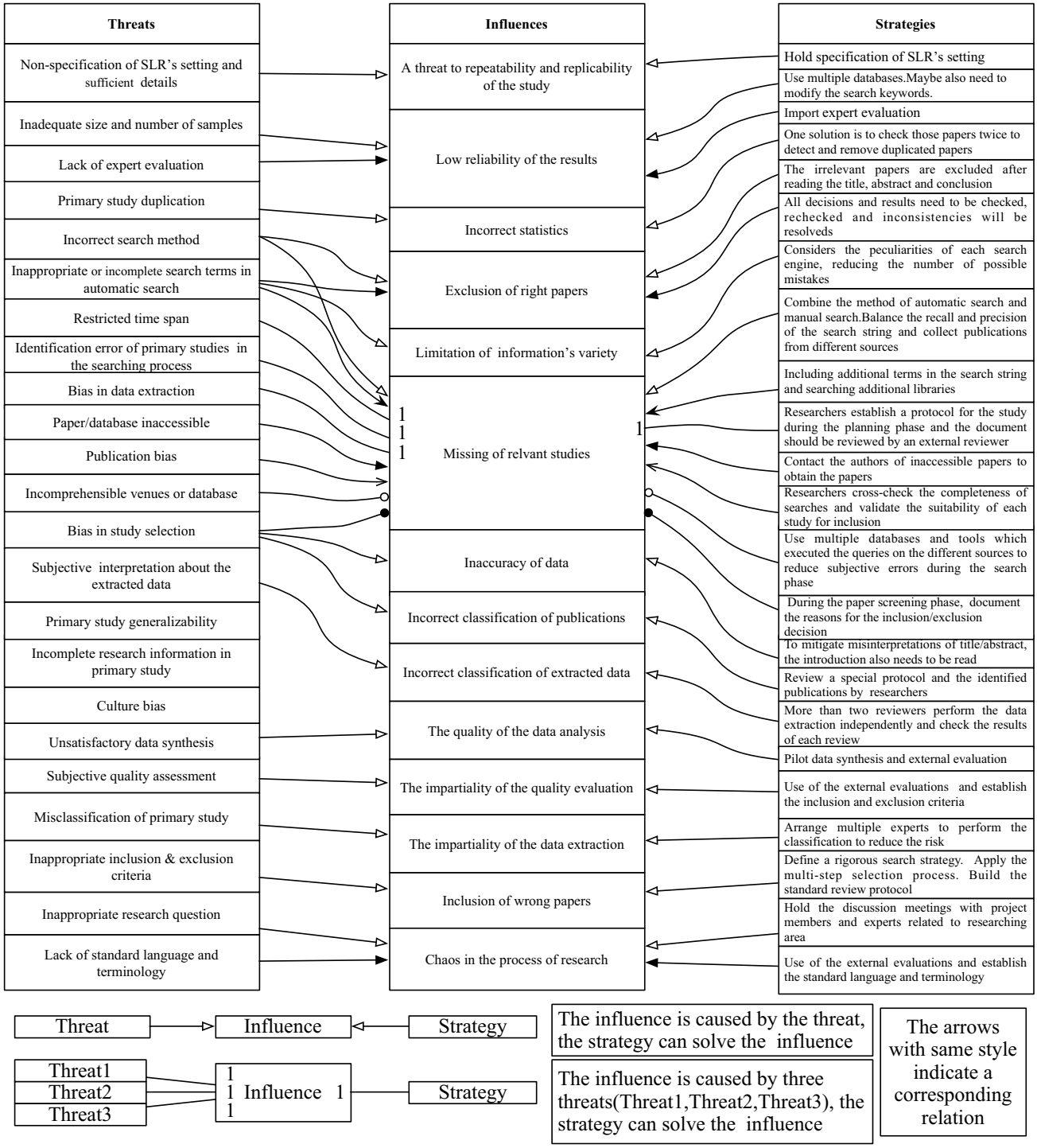

Figure 2. Threats & Influences & Strategies Triadic Relations

Inappropriate inclusion & exclusion criteria: 1) Define a rigorous search strategy [S11]; 2) Apply the multi-steps selection process; 3) Build the standard review protocol [S36].

Inappropriate research questions: Hold the discussion meetings with project members and experts related to the research area [S7].

Bias in Study Selection: 1) You must carefully read through the remaining papers to mitigate misinterpretation. It will help to mitigate misinterpretations of title/abstract [S8]. 2) Develop detailed guidelines in the review protocol prior to the start of the review. During the paper-screening phase, document the reasons for the inclusion/exclusion

decision [S10]. 3) Review a standard protocol and the identified publications by researchers [S14].

Paper/database inaccessible: Contact the authors of inaccessible papers to obtain the papers.

Misclassification of primary study: Arrange multiple experts to perform the classification to reduce the risk [S2] [S4] [S20].

Primary study duplication: One solution is to check those papers twice to detect and remove duplicated papers.

Publication bias: 1) Researchers cross-check the completeness of the search and validate the suitability of each study for inclusion [S5]. 2) Researchers can search the gray literature [S8].

Bias in Data Extraction: 1) Researchers establish a protocol for the study during the planning phase and the document should be reviewed by an external reviewer [S6]. 2) More than two reviewers perform the manual search independently and check the results of each review [S14].

Unsatisfactory Data Synthesis: Pilot data synthesis and external evaluation [S17].

Subjective quality assessment: 1) Use the external evaluation [S17]. 2) Establish the inclusion and exclusion criteria [S9].

We put forward some strategies to help cope with the TTVs in SLRs. 1) Establish the protocol for the study during the planning phase, which is reviewed by an external reviewer. 2) Well-defined inclusion and exclusion criteria are helpful to reduce the identification error of primary studies. 3) All decisions and results are double-checked by at least one other person. 4) Design a search method combining the method of automatic search, manual search and snowballing. 5) Use multiple databases and tools, which execute the queries on the different sources to reduce human errors during the search phase. 6) Check those papers twice to detect and remove duplicated papers.

IV. Related Work

SLR is a relatively new research method in SE. Some experience reports talk about the TTVs, and we make a comparison of their work and ours. Ali Babar and Zhang started an empirical research program that aims to contribute to the growing body of knowledge about systematic reviews in software engineering [6]. Zhang and Ali Babar investigated the adoption and use of SLRs in SE research from various perspectives using multi-method approach [7]. Their findings provide interesting insights into different aspects of SLRs.

Lack of expert judgment may lead to erroneous conclusions. Some experience reports put forward the lack of expert evaluation is one of the threats, but we found it is not obvious in our review.

Imtiaz et al. [11] carried out a tertiary study about the conducting SLRs experience in software engineering. The experience gathered from the included evidence highlights search strategies as the most problematic activity of SLR. Data extraction and inclusion & exclusion activities also have many problems. We usually use experience to effectively handle them. Some of the reported problems in this phase are difficulty to select keywords, identify representative synonyms and alternative terms, limitations of online databases, ensuring coverage of relevant material etc. The results show the evidence supports our conclusion.

In some experience reports, researchers explicitly discuss begin to focus on relevant threats to validity of empirical research is general and corresponding strategies against them.

Biffl et al. introduced a TTV knowledge base (KB) to support experiment planners in identifying relevant TTVs and control them [9]. By conducting controlled experiments, the feasibility of using the TTV KB is verified. With a focus on the tradeoff between internal validity, external validity and replication, Siegmund et al. asked the community how empirical research should take place in software engineering and complemented with a literature review [10]. They found that the opinions differ considerably, and that there is no consensus in the community when to focus on internal or external validity and how to conduct and review replications.

A relatively new experience report was conducted by Imtiaz et al. [11]. This research recorded the reported experiences of conducting SLR, that why benefit new researchers. It documented reported experiences in each phase. The main problem they mentioned is how to carefully select keywords, synonyms and alternative terms.

V. Threats To Validity

Construct Validity: Our research questions may not be able to completely cover all the SLRs that describe in the related content.

Internal Validity: In order to exhaustively identify SLRs with TTVs component and ensure that the process of papers' selection was unbiased as far as possible, the approach of quasi-gold standard (QGS) [4][8] was adopted, which systematically integrates manual and automated search strategies and suggests a relatively rigorous approach for search performance evaluation in terms of sensitivity and precision. Although we only searched four online digital libraries, they are believed to cover the majority of the high quality publications in SE. To capture as many SLRs as possible, however, we also used the snowballing as the complementary search to reduce the possibility of missing relevant SLRs. In addition, the search strategy was developed by the three research students and reviewed by the supervisor.

External Validity: We selected SLRs that include a discussion about TTVs from 2004 to the first half of 2015. The excluded SLRs without TTV issues reported may affect the generalizability of our result.

Conclusion Validity: We extracted the data from the selected SLRs with TTVs discussions, the influences and the strategies to cope with the threats. To ensure the correctness of the extracted data, the protocol was developed to define the data extraction strategy and format.

The review protocol was proposed by the three authors, and was then reviewed by their supervisor. We defined a data extraction form to obtain consistent extraction of relevant information and checked whether the data to be extracted would address the research questions. Moreover, the cross-check was necessary among the reviewers, and again we had at least two researchers extracting data independently. The supervisor dealt with any divergences and disagreements during the process.

VI. Conclusion And Future Work

Since TTVs may decrease the credibility of the conclusion of an SLR, this paper reported and discussed TTVs in SLRs in SE. Our aim is to provide researchers with a set of common threats to identify and deal with TTVs for SLRs in SE. Based on our report, researchers could easily identify the potential threats of their study and could avoid these threats during the planning phase of the review.

We also thoroughly analyzed the reported influences of threats of the included SLRs. We find the most reported influence of TTVs is missing relevant papers. Since the most common threat is caused by the bias of researchers, pilot of exact phase of review or cross-check could be an effective factice against most threats.

We collected information about the strategies to corresponding influences caused by TTVs. Researchers are able to choose appropriate strategies to minimize the influences or cope with the TTVs. Since the most common threat is caused by the bias of researchers, pilot review or cross-check could be an effective strategy against for most threats.

Based on our observation, researchers pay greatly attention to the internal and conclusion validity. The construct and external validity are also important in SLRs, but often ignored. These two types of validity need more considerations in the future SLR researches.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.61572251).

References

[1] B. A. Kitchenham, T. Dyba and M. Jorgensen, “Evidence-based software engineering,” in *Proceedings of the 26th International Conference on Software Engineering (ICSE’04)*, Washington, DC, USA, IEEE Computer Society, May 2004, pp. 273–281.

[2] B. A. Kitchenham and S. M. Charters, “Guidelines for performing systematic literature reviews in software engineering,” Technical Report, Keele University and University of Durham, 2007.

[3] H. Zhang and M. A. Babar, “Systematic reviews in software engineering: An empirical investigation,” *Information and Software Technology*, vol. 55, no. 7, pp. 1341 – 1354, Jul. 2013.

[4] H. Zhang, M. A. Babar, and P. Tell, “Identifying relevant studies in software engineering,” *Information and Software Technology*, vol. 53, no. 6, pp. 625 – 637, Jun. 2011.

[5] Y. Zhou, H. Zhang, X. Huang, S. Yang, M. Ali Babar, and H. Tang, “Quality assessment of systematic reviews in software engineering: A tertiary study,” in *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (EASE’15)*, Nanjing, China, ACM, Apr. 2015, pp. 1–14.

[6] M. Ali Babar and H. Zhang, “Systematic literature reviews in software engineering: Preliminary results from interviews with researchers,” in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement (ESEM’09)*, Lake Buena Vista, Florida, USA, IEEE Computer Society, Oct. 2009, pp.346–355.

[7] H. Zhang and M. Ali Babar, “An empirical investigation of systematic reviews in software engineering,” in *2011 International Symposium on Empirical Software Engineering and Measurement (ESEM’11)*, Banff, AB, Canada, IEEE Computer Society, Sept. 2011, pp. 87–96.

[8] H. Zhang and M. Ali Babar, “On searching relevant studies in software engineering,” in *Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering (EASE’10)*, Swinton, UK, British Computer Society, Apr. 2010, pp. 111–120.

[9] S. Biffl, M. Kalinowski, F. Ekaputra, A. A. Neto, T. Conte, and D. Winkler, “Towards a semantic knowledge base on threats to validity and control actions in controlled experiments,” in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM’14)*, Torino, Italy, ACM, 2014, pp.1–4.

[10] J. Siegmund, N. Siegmund, and S. Apel, “Views on internal and external validity in empirical software engineering,” in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE'15)*, IEEE, May 2015, pp. 9–19.

[11] S. Imtiaz, M. Bano, N. Ikram, and M. Niazi, “A tertiary study: Experiences of conducting systematic literature reviews in software engineering,” in *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering (EASE’13)*, Porto de Galinhas, Brazil, ACM, Jan. 2013, pp.177–182.

[12] J. Biolchini, P. G. Mian, A. Candida and C. Natali, “Systematic review in software engineering,” Technical Report, Systems Engineering and Computer Science Department, COPPE/UFRJ, 2005.

[13] T. Dyba, B. A. Kitchenham, and M. Jorgensen, “Evidence based software engineering for practitioners,” *IEEE Software*, vol. 22, no. 1, pp. 58–65, Jan. 2005.

[14] B. Kitchenham, “Procedures for performing systematic reviews,” Keele University Technical Report, Jul. 2004.

[15] S. Montagud and S. Abrahao, “Gathering current knowledge about quality evaluation in software product lines,” in *Proceedings of the 13th International Software Product Line Conference (SPLC’09)*, San Francisco, California, USA, Carnegie Mellon University, Aug. 2009,pp. 91–100.

[16] A. M. Fernández-Sáez, M. Genero, and M. R. Chaudron, “Empirical studies concerning the maintenance of uml diagrams and their use in the maintenance of code: A systematic mapping study,” *Information and Software Technology*, vol. 55, no. 7, pp. 1119 – 1142, Jul. 2013.

[17] C. Pacheco and I. Garcia, “A systematic literature review of stakeholder identification methods in requirements elicitation,” *Journal of Systems and Software*, vol. 85, no. 9, pp. 2171 – 2181, Sept. 2012.

[18] S. Heckman and L. Williams, “A systematic literature review of actionable alert identification techniques for automated static code analysis,” *Information and Software Technology*, vol. 53, no. 4, pp. 363 – 387, Apr. 2011.

[19] H. K. Wright, M. Kim, and D. E. Perry, “Validity concerns in software engineering research,” in *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research (FoSER’10)*, Santa Fe, New Mexico, USA, ACM, Nov. 2010, pp. 411–414.

[20] K. Petersen, “Measuring and predicting software productivity: A systematic map and review,” *Information and Software Technology*, vol. 53, no. 4, pp. 317 – 343, 2011.

[21] D. S. Cruzes and T. Dyba, “Recommended steps for thematic synthesis in software engineering,” in *2011 International Symposium on Empirical Software Engineering and Measurement (ESEM’11)*, IEEE, Sept. 2011, pp. 275–284.

[22] G. W. Noblit and R. D. Hare, Meta-ethnography: Synthesizing qualitative studies. sage, 1988.