Identifying, categorizing and mitigating threats to validity in software engineering secondary studies

Ampatzoglou, Apostolos; Bibi, Stamatia; Avgeriou, Paris; Verbeek, Marijn; Chatzigeorgiou, Alexander

Published in:
Information and Software Technology

DOI:
10.1016/j.infsof.2018.10.006

Publication date:
2019

Information and Software Technology 106 (2019) 201–230

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Identifying, categorizing and mitigating threats to validity in software engineering secondary studies

Apostolos Ampatzoglou[a,*], Stamatia Bibi[b], Paris Avgeriou[c], Marijn Verbeek[c], Alexander Chatzigeorgiou[a]

[a] *Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece*
[b] *Department of Informatics and Telecommunications, University of Western Macedonia, Kozani, Greece*
[c] *Department of Mathematics and Computer Science, University of Groningen, the Netherlands*

ARTICLE INFO

Keywords:
Empirical software engineering
Secondary studies
Threats to Validity
Literature Review

ABSTRACT

Context: Secondary studies are vulnerable to threats to validity. Although, mitigating these threats is crucial for the credibility of these studies, we currently lack a systematic approach to identify, categorize and mitigate threats to validity for secondary studies.
Objective: In this paper, we review the corpus of secondary studies, with the aim to identify: (a) the trend of reporting threats to validity, (b) the most common threats to validity and corresponding mitigation actions, and (c) possible categories in which threats to validity can be classified.
Method: To achieve this goal we employ the tertiary study research method that is used for synthesizing knowledge from existing secondary studies. In particular, we collected data from more than 100 studies, published until December 2016 in top quality software engineering venues (both journals and conference).
Results: Our results suggest that in recent years, secondary studies are more likely to report their threats to validity. However, the presentation of such threats is rather ad hoc, e.g., the same threat may be presented with a different name, or under a different category. To alleviate this problem, we propose a classification schema for reporting threats to validity and possible mitigation actions. Both the classification of threats and the associated mitigation actions have been validated by an empirical study, i.e., Delphi rounds with experts.
Conclusion: Based on the proposed schema, we provide a checklist, which authors of secondary studies can use for identifying and categorizing threats to validity and corresponding mitigation actions, while readers of secondary studies can use the checklist for assessing the validity of the reported results.

1. Introduction

Empirical Software Engineering (ESE) research focuses on the application of empirical methods on any phase of the software development lifecycle. The three predominant types of empirical research are [44,47]: (a) surveys, which are performed through questionnaires or interviews on a sample in order to obtain characteristics of a population [36]; (b) *case studies*, which study phenomena in a “real-world” context, especially when the boundaries between phenomenon and context are not clear [51]; and (c) *experiments*, which have a limited scope and are most often run in a laboratory setting, with a high level of control [47]. During the last years and mainly due to the rise of the Evidence-Based Software Engineering (EBSE) Paradigm[1] [22], two other types of studies have become quite popular [15]:

- **Systematic Literature Reviews** (SLRs) use data from previously published studies for the purpose of *research synthesis*, which is the collective term for a family of methods for summarizing, integrating and, when possible, combining the findings of different studies on a topic or research question. Such synthesis can also identify crucial areas and questions that have not been addressed adequately with past empirical research. It is built upon the observation that no matter how well-designed and executed, empirical findings from individual studies are limited in the extent to which they may be generalized [18].

[1] EBSE is a movement in the software engineering research that *aims to provide the means by which current best evidence from research can be integrated with practical experience* [22].

[*] Corresponding author.
E-mail address: apostolos.ampatzoglou@gmail.com (A. Ampatzoglou).

https://doi.org/10.1016/j.infsof.2018.10.006
Received 20 February 2018; Received in revised form 4 October 2018; Accepted 6 October 2018
Available online 9 October 2018

- **Systematic Mapping Studies** which use the same basic methodology as SLRs but aim to identify and classify all research related to a broad software engineering topic rather than answering questions about the relative merits of competing technologies that conventional SLRs address. They are intended to provide an overview of a topic area and identify whether there are sub-topics with sufficient primary studies to conduct conventional SLRs and also to identify sub-topics where more primary studies are needed [21].

The strength of evidence produced by ESE research depends largely on the use of systematic, rigorous guidelines on how to conduct, and report empirical results (see e.g., for experiments [47], for SLRs [18], for mapping studies [34], for surveys [36], and for case studies [38]). One of the most crucial parts of conducting an empirical study is the management of threats to validity, i.e., possible aspects of the research design that in some way compromise the credibility of results. Despite this crucial role, we currently lack guidelines on how to identify, mitigate, and categorize threats to validity in secondary studies; this is in contrast to experiments, case studies and surveys, where mature guidelines exist. Due to this reason, researchers either do not report threats to validity for secondary studies, or report them in an ad hoc way (see Section 5). Specifically, the most common issues found in practice, concern threats to validity being:

- **Completely missing** from certain studies. Thus, such studies do not provide any mitigation actions for them;
- **Incorrectly categorized**. The same threat is classified in different categories by different researchers (e.g., *study selection bias* is categorized in some studies as threat to *internal* and in others as a threat to *conclusion validity*. Also, in some cases threats are **inefficiently categorized** based on guidelines for other types of empirical research (e.g., for experiments [45], or for case studies [38]), or under a custom categorization, which is ***not uniform***. One possible reason for this problem is the fact that threat categories are not orthogonal, especially in cases where they stem from different schools of thought or guidelines (see Section 2.1). For example, reliability examines if the results of a study depend highly on the involved researchers. In turn, this relates to conclusion validity, in the sense that people are prone to biases (e.g. due to previous experiences, preferences on research, etc.);
- **Inconsistently named**. The same threat is reported with a different name by different researchers (e.g., the terms *publication bias* and *researcher bias* are used for describing the same threats);
- **Inconsistently mitigated**. The same threat is mitigated differently by different researchers. Although this provides a variety of available mitigation actions, some mitigation actions are ineffective and cause confusion to readers who consider following them.

These issues, in turn lead to a difficulty in evaluating the validity of the reported results and hinder a uniform comparison between secondary studies. In addition, the lack of guidance for mitigating threats to validity, which could serve as a reference point, makes it more difficult to reuse mitigation strategies, as well as to consistently identify and categorize both threats and mitigation actions.

To address this problem, we conducted a tertiary study (i.e., an SLR on secondary studies), so as to retrieve and analyze how software engineering secondary studies identify, categorize and mitigate threats to validity. The objective of this tertiary study is: "to summarize *secondary studies that report threats to validity*, with the aim of identifying: (a) the *frequency of reporting* threats to validity over the years, (b) the most *common threats to validity* and (c) the corresponding *mitigation actions*, and (d) a possible *classification schema of threats to validity*". The main outcomes of the study are a classification schema for threats to validity and a checklist that can be used while conducting/evaluating secondary studies. The outcomes are expected ***to contribute towards establishing a standard and consistent way of identifying, categorizing and mitigating threats to validity of secondary studies.*** In addition to that, in order to enrich the outcomes of this work we explored existing literature in two related research sub-fields: (a) secondary studies in medical science (i.e., the area from where the Evidence-Based paradigm has emerged from), and (b) guidelines for conducting secondary studies. Related studies from medical science and the guidelines for performing secondary studies has led to the identification of best practices in secondary studies that can be applied as mitigation actions for minimizing of effects of a validity threat, enriching the provided checklist that has been derived from the classification schema. Finally, acknowledging the subjectivity in the qualitative nature of this work, we validated the outcomes through a Delphi method based on the opinion of experts in secondary studies and empirical studies in general. The Delphi method was iterated in three rounds and provided preliminary evidence for the merits of the classification schema and checklist.

We note that literature reviews have been performed long before the advent of the terms 'Systematic Mapping Study' and 'Systematic Literature Review' and corresponding guidelines. We also acknowledge that secondary studies can be performed without following the guidelines of SMSs and SLRs (especially before the two terms become popular). However, such non-systematic literature reviews have not reported (in the vast majority of the cases) threats to their conclusions. Reporting of threats became popular once specific guidelines were proposed and adopted in the context of the EBSE paradigm. Thus, for a study aiming at systematically analyzing the reported threats, we consider it proper to focus on the studies that have adopted the corresponding guidelines. For the rest of the study, when we refer to secondary studies, we refer to Systematic Mapping Studies and Systematic Literature Reviews.

The rest of the paper is organized as follows: Section 2 presents related work, i.e., categories of threats to validity in other empirical methods; Section 3 presents our tertiary study protocol; Section 4 reports on the results; and Section 5 discusses the proposed guidelines for identifying, categorizing and mitigating threats to validity for secondary studies in software engineering. In Section 6, we present the design and results of our validation study, whereas in Sections 7 and 8 we present threats to validity and conclude the paper.

2. Related work

The empirical software engineering literature points out the relevance and importance of identifying and recording validity threats, as an aspect of research quality [12,32] and [35]. According to Perry et al. [32] the structure of an empirical study in SE should include a section of threats to validity. This section should discuss the influences that may limit the authors' and readers' ability to interpret or draw conclusions from the study's data. In addition, Jedlitschka et al. [17] suggest that each controlled experiment in SE should have a subsection named "*Limitation of the study*" where all threats that may have an impact on the validity of results should be mentioned. Furthermore, Kitchenham [22] has also underlined the importance of threats to validity, by highlighting that the implications of a validity threat should be addressed and thoroughly discussed. Finally, Sjoberg et al. [42] emphasize the scope of validity of the results of a SE study; the term 'scope of validity' is interpreted as the population of actors, technologies, activities, software systems for which the results of a study are valid. The scope of validity is considered to be crucial for producing general knowledge synthesized by comparing and integrating results from different studies.

In this section we present related work, under three perspectives. First, we present how threats to validity are categorized in the empirical software engineering field (see Section 2.1). Second, in Section 2.2, we present studies that are related to the identification and reporting of threats to validity in medical science. This can provide valuable input for our work, since medical research is considered a more mature field in secondary study design and execution and has already inspired the guidelines for conducting secondary studies in software engineering. Finally, in Section 2.3, we present the most common guidelines for per-

Table 1
Categories of Threats to Validity in ESE Research.

Conclusion validity: Originally called "statistical conclusion validity", this aspect deals with the degree to which conclusions reached (e.g. about relationships between factors) are reasonable within the data collected. Researcher bias, for example, can greatly impact conclusions reached and can be considered to be a threat to conclusion validity. Similarly, statistical analysis may lead to weak results that can be interpreted in different ways according to the bias of the researcher. In either case the researcher may reach the wrong conclusion [47].

Reliability: This aspect is concerned with to what extent the data and the analysis are dependent on the specific researchers. Example of this type of threat is the unclear coding of collected data. If a researcher produces certain results, then, other researchers should be able to reproduce identical results following the same methodology of the study [38].

Internal validity: This aspect relates to the examination of causal relations. Internal validity examines whether an experimental treatment/condition makes a difference or not, and whether there is evidence to support the claim [47].

Construct validity: Defines how effectively a test or experiment measures up to its claims. This aspect deals with whether or not the researcher measures what is intended to be measured [47].

External validity: The concern of this aspect is whether the results can be generalized. During the analysis of this validity, the researcher attempts to see if findings of the study are of relevance for others. In the case of quantitative research (experiments), this primarily relies on the chosen sample size. In contrast, case studies have normally a low sample size, so the researcher has to try and analyze to what extent the findings can be related to other cases [47].

forming secondary studies in the software engineering domain, as they can also provide input for our work.

2.1. Threats to validity in empirical software engineering

Threats to validity have been often categorized in the literature of general research methods in different types. Initially, Cook and Campbell [8][2] recorded four types of validity threats in quantitative experimental analysis: *statistical conclusion validity, internal validity, construct validity of putative causes and effects* and *external validity*. Concerning qualitative research, Maxwell [29] provided a general categorization of threats that can be mapped to Cook and Campbell's categorization as follows: *theoretical validity* (construct validity), *generalizability* (internal, external validity), and *interpretive validity* (statistical conclusion validity). An additional threat category, mentioned by Maxwell [29], is *descriptive validity,* which is relevant only for qualitative studies. Descriptive validity reflects the accuracy and objectivity of the information gathered. For example, when researchers collect statements from participants, threats to validity can be related to the way that researchers recorded or transcribed the statements. Other types of validity threats that are found in literature are: reliability [38,51], transferability, credibility and confirmability [27], uncontrollability, and contingency [14].

In the empirical SE community there are two main schools on reporting threats to validity: (a) Wohlin et al. [47] who adopted Cook and Campbell's [8] categorization of validity threats and presented four main types of threats to validity for quantitative research within software engineering: ***conclusion, internal, construct***, and ***external*** validity; and (b) Runeson et al. [38] who discussed four main types of validity threats for case studies within software engineering: ***reliability, internal, construct***, and ***external*** validity. The threats of Runeson et al. [38] are similar to those of Wohlin et al. [47] with the exception of reliability replacing conclusion validity.

Biffl et al. [4] argue that researchers should also consolidate actual experimental research on a specific topic to complement existing generic threats and guidelines when performing their research. The tradeoff between internal and external validity has been addressed by Siegmund et al. [40], where the authors performed a survey and concluded that externally valid papers are of greater practicality while internally valid studies seem to be unrealistic. Additionally, the study examined the impact of replication studies and found that although researchers realize the necessity of such studies they are reluctant to conduct or review them mainly due to the fact that there are no guidelines for performing them [40]. A list of definitions of the union of the aforementioned categories of threats to validity (i.e. from [38] and [47]) are presented in Table 1.

Petersen et al. [35] based on the categorizations of threats to validity suggested by Maxwell, suggested a check list that can help researchers identify the threats applicable to the type of research performed by reporting first their world-view and then the research method applied. A secondary study attempting to assess the practices in reporting validity threats in ESE [12] concluded that more than 20% of the studied papers contain no discussion of validity threats and the ones that do discuss validity threats on average contain 5.44 threats.

Regarding threats to validity for secondary studies in software engineering, we have been able to identify only one related work. In particular, Zhou et al. [53] have performed a tertiary study on more than 300 secondary studies until 2015. The authors have identified 23 threats to validity for secondary studies, and organize the consequences of these studies into four categories: internal, external, conclusion, and construct validity. To alleviate these threats the authors maps the threats and possible consequences to 24 mitigation strategies. This paper shares common goals with our study, however, ours is broader in the sense that: (a) it covers a wider timeframe (until 2017 instead of middle of 2015); (b) it focuses only on top-quality venues, which are expected to pay special attention in the proper application of methodological guidelines, such as the proper reporting of threats to validity, a fact that increases the quality of the obtained data; and most importantly (c) our study answers two additional RQs, providing a classification schema and a checklist for identifying, mitigating, and reporting threats to validity. In addition to this, as indirect related work (especially in terms of mitigation actions), in Section 2.3 we present a review of guidelines on secondary studies in software engineering.

2.2. Threats to validity in medical science

In this section we report on quality assessment strategies for systematic reviews from medical science literature. While there is no classification of threats to validity for secondary studies or corresponding mitigation actions in medical research, these quality assessment strategies can provide useful input for deriving such outcomes in the software engineering domain. Particularly we identify a number of quality assessment criteria based on the guidelines, the checklists and protocols found in medical research literature. These quality assessment criteria are subsequently classified into five categories, presented in Table 2, based on the aspect that they address: (a) primary study selection process, (b) validity of primary studies (c) data reliability, (d) research design and (e) reporting process. An additional factor that affects the quality of secondary studies is the level of detail and completeness of reporting. The criteria in Table 2 will be exploited after the development of the proposed classification schema. In particular, we check if the criteria in Table 2 are included in the list of mitigation actions; if not we incorporate them in the proposed checklist, as best practices (see Section 5).

The methodological quality of experiments and reviews performed in the medical domain was assessed by Downs et al. [10] who formed a checklist consisting of 26 items/ questions for assessing the quality of randomized and non-randomized healthcare studies. The main quality aspects captured in this checklist involved the Reporting stage, the External Validity, the Internal Validity and the Selection Bias. Further-

[2] Before publishing this paper (i.e., [8]), Cook and Campbell had published an online chapter focused on Conclusion and Internal validity threats.

Table 2
Quality Assessment Criteria in Medical Studies.

Primary study selection:
Was there duplicate study selection and data extraction? [31,39]
Was a comprehensive literature search performed? [7,30,31,39,43]
Was the status of publication (i.e. grey literature) used as an inclusion criterion? [39]
Have additional studies been identified? [52]
Assessing Validity of Primary Studies:
Were the eligibility criteria specified? [45]
Were statistical results and measures of variability presented for the primary outcome measures? [1,10,30,45]
Was the quality of the included studies assessed? [16,31,39,45,52]
Data reliability:
Was the likelihood of publication bias assessed? [11,37,39]
Were methods for data extraction and analysis evaluated? [10,30,31,39,52]
Was there any conflict of interest stated? [39]
Research Design:
Was an ’a priori’ design provided? [31,39,43]
Was the scientific quality of the included studies used appropriately in formulating conclusions? [39,43]
Is a database, containing the relevant data, available as a resource for intervention planners and researchers? [52]
Was other pertinent information identified to ensure study intervention’s applicability in settings and populations other than that studied by the investigators? [52]
Reporting Process:
Was a list of studies (included and excluded) provided? [31,39]
Were the characteristics of the included studies provided? [39,52]
Was the scientific quality of the included studies documented? [7,39]

more, the Prisma-P meta-analysis protocol for systematic reviews has been proposed by Moher et al. [31] consisting of a checklist of 17 items categorized into three main sections: Administrative information, Introduction and Methods. The Administrative section represents mainly initial information on the authors, the funding and the title of the study, the Introduction section includes details on the rationale and the objectives of the study while the Methods section specifies the information sources, the study selection criteria, the search string and the data analysis methods employed within the scope of the meta-analysis study. Moreover, the medical domain uses the Cohraine database[3] (including the Database of Abstracts of Reviews of Effects) [7] that contains more than 15,000 abstracts of high quality reviews that are independently appraised by two reviewers according to the following six criteria: reporting of inclusion/exclusion criteria, adequacy of search, data synthesis, validity assessment of primary studies included and detailed presentation of individual studies referenced.

Shea et al. [39] developed an instrument to assess the methodological quality of systematic reviews building upon previous tools, empirical evidence and expert consensus. The tool was based on 11 components that summarized and synthesized evidence from the initial quality checklist that included 37 items. These items were subjected to principal component analysis, and Varimax rotations. The validity of systematic reviews is also assessed by Slocum et al. [43] who advise the researchers of review studies to carefully define research questions and focus on them, and to systematically search the literature, validate primary studies and document the search process so as to enable reproducibility. Furthermore, publication bias is acknowledged as a significant problem by Dwan et al. [11] as it produces outcome reporting bias, due to the fact that positive results are easier to publish. In that case the authors advise the researchers to improve the reporting of trials (primary studies). Publication bias is also addressed by Rothstein [37] who suggests the use of funnel plots to detect it and the use of cumulative meta-analysis to assess its impact.

Verhaegen et al. [45] adopted the Delphi technique, as a consensus method, to identify quality criteria for selecting the primary studies (referred to as Medical Clinical Trials) that participate in healthcare literature reviews. A three-round Delphi was performed where each participant answered questions in the form of “Should this item be included into the criteria list?” utilizing a 5-point Likert scale. The quality criteria derived from the final Delphi round are included in Table 2. We note that we isolated the criteria that are not specialized in medical research. In this context, blind assessment of clinical trial studies, treated as primary studies in medical reviews, was proposed in [16]. The findings of [16] suggest that blind assessments are reliable producing more consistent scores compared to open assessments. Furthermore, a data collection instrument for performing systematic reviews for disease preventions was proposed by Zaza et al. [52]. The authors concluded in a six point assessment form. The content of the form was developed by reviewing methodologies from other systematic reviews; reporting standards established by major health and social science journals; the evaluation, statistical and meta-analytic literature; and by soliciting expert opinion. Avellar et al. [1] scanned 19 reviews in the medical field in order to examine the level to which external validity is addressed. The results revealed that most studies lack statistical representativeness in terms of the generalizability threat and focus only on factors likely to increase the heterogeneity of primary studies and context [1]. With respect to these results Avellar et al. [1] split external validity into three aspects: generalizability (related to the number of studies reporting the same result and the settings required to achieve a certain result), applicability (demographics of the population in which a certain result is achieved) and feasibility (description of an intervention required to be performed, in medical studies it is related to the dosage, the staff training, the cost).

2.3. Overview of guidelines for conducting secondary studies in software engineering

In this section we present the most common guidelines for performing secondary studies in the software engineering domain, in an attempt to consider relevant methodological problems and gain insights from the reported advice and lessons learned. A summary of the guidelines provided for conducting secondary studies in the software engineering field is presented in Fig. 1. Similarly to the case of the quality assessment criteria in medical studies, we intend to use these guidelines after the development of the proposed classification schema. In particular, we check if the practices reported in Fig. 1 are included in the list of mitigation actions of the classification schema. Those that are not, will be incorporated in the proposed checklist, as best practices.

The guidelines of Kitchenham et al. [18] are considered seminal for performing Systematic Literature Reviews (SLRs) in software engineering. Three major stages for performing SLRs are suggested: Planning, Conducting and Reporting, each of which including several mandatory

[3] http://community.cochrane.org/editorial-and-publishing-policy-resource/overview-cochrane-library-and-related-content/databases-included-cochrane-library/database-abstracts-reviews-effects-dare

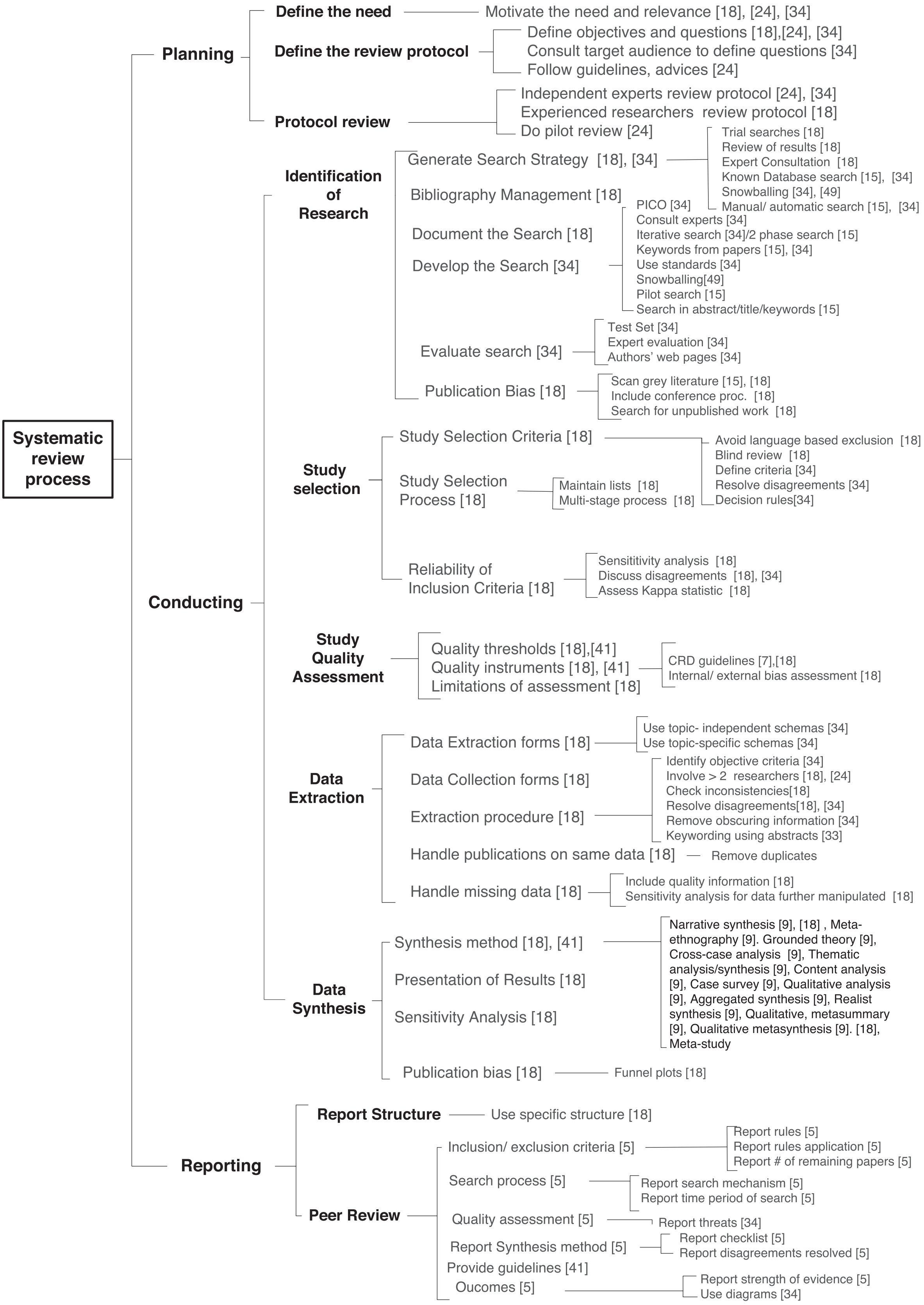
Systematic review process
Planning
Define the need
Motivate the need and relevance [18], [24], [34]
Define the review protocol
Define objectives and questions [18],[24], [34]
Consult target audience to define questions [34]
Follow guidelines, advices [24]
Protocol review
Independent experts review protocol [24], [34]
Experienced researchers review protocol [18]
Do pilot review [24]
Conducting
Identification of Research
Generate Search Strategy [18], [34]
Trial searches [18]
Review of results [18]
Expert Consultation [18]
Known Database search [15], [34]
Snowballing [34], [49]
Manual/ automatic search [15], [34]
Bibliography Management [18]
Document the Search [18]
Develop the Search [34]
PICO [34]
Consult experts [34]
Iterative search [34]/2 phase search [15]
Keywords from papers [15], [34]
Use standards [34]
Snowballing[49]
Pilot search [15]
Search in abstract/title/keywords [15]
Evaluate search [34]
Test Set [34]
Expert evaluation [34]
Authors' web pages [34]
Publication Bias [18]
Scan grey literature [15], [18]
Include conference proc. [18]
Search for unpublished work [18]
Study selection
Study Selection Criteria [18]
Avoid language based exclusion [18]
Blind review [18]
Define criteria [34]
Resolve disagreements [34]
Decision rules[34]
Study Selection Process [18]
Maintain lists [18]
Multi-stage process [18]
Reliability of Inclusion Criteria [18]
Sensititivity analysis [18]
Discuss disagreements [18], [34]
Assess Kappa statistic [18]
Study Quality Assessment
Quality thresholds [18],[41]
Quality instruments [18], [41]
Limitations of assessment [18]
CRD guidelines [7],[18]
Internal/ external bias assessment [18]
Data Extraction
Data Extraction forms [18]
Use topic- independent schemas [34]
Use topic-specific schemas [34]
Data Collection forms [18]
Extraction procedure [18]
Identify objective criteria [34]
Involve > 2 researchers [18], [24]
Check inconsistencies[18]
Resolve disagreements[18], [34]
Remove obscuring information [34]
Keywording using abstracts [33]
Handle publications on same data [18]
Remove duplicates
Handle missing data [18]
Include quality information [18]
Sensitivity analysis for data further manipulated [18]
Data Synthesis
Synthesis method [18], [41]
Narrative synthesis [9], [18] , Meta-ethnography [9]. Grounded theory [9], Cross-case analysis [9], Thematic analysis/synthesis [9], Content analysis [9], Case survey [9], Qualitative analysis [9], Aggregated synthesis [9], Realist synthesis [9], Qualitative, metasummary [9], Qualitative metasynthesis [9]. [18], Meta-study
Presentation of Results [18]
Sensitivity Analysis [18]
Publication bias [18]
Funnel plots [18]
Reporting
Report Structure
Use specific structure [18]
Peer Review
Inclusion/ exclusion criteria [5]
Report rules [5]
Report rules application [5]
Report # of remaining papers [5]
Search process [5]
Report search mechanism [5]
Report time period of search [5]
Quality assessment [5]
Report threats [34]
Report Synthesis method [5]
Report checklist [5]
Report disagreements resolved [5]
Provide guidelines [41]
Oucomes [5]
Report strength of evidence [5]
Use diagrams [34]

Fig. 1. Overview of guidelines for performing secondary studies.

activities. A detailed and updated guide on performing systematic reviews can be found in the study by Kitchenham et al. [25] where all the stages and the corresponding activities are further analyzed. Similarly, Petersen et al. [33] provided guidelines for performing SMSs in software engineering, following a five-stage process that includes, research question identification, conducting the search, screening of papers, keywording using abstracts, and data extraction and mapping. This process of performing SMSs was updated by Petersen et al. [34].

According to Budgen et al. [5] the reporting process of secondary studies is very crucial and should provide details about the inclusion/exclusion criteria of primary studies, the search process adopted for the retrieval of primary studies, the quality assessment of the review process, the data synthesis methodology and the clear reporting of outcomes. Similarly, Cruzes and Dyba [9] emphasized the data analysis stage, during the execution of secondary studies, providing a list of data synthesis methods with the corresponding description. They reached the conclusion that only 50% of the examined secondary studies performed data synthesis. Regarding the searching stage, Wohlin explored the snowballing approach as an alternative method for the primary study identification stage [49].

Among the most common problems related to secondary study research, as identified by Kitchenham et al. [24], is the difficulty to perform complex automated searches in the digital libraries, the time and effort required to complete the study, the definition of the research protocol and the quality assessment of the primary studies. Kitchenham et al. [24] advise the authors of secondary studies to follow reported guidelines (such as those discussed in Fig. 1), clearly define research questions, validate externally the research protocol and work in pairs so that one author extracts data and the other one performs checks. The results of Wholin et al. [49] point out that snowballing can complement traditional database search method. Another problem regarding the process of conducting secondary studies is that the majority of the SLRs does not address the quality of primary studies and fail to provide guidelines for practitioners [41]. Imtiaz et al. [15] analyzed the findings of 116 secondary studies performed in the field of software engineering and reported that the Search Strategy, the Online Databases and the Planning and Data Extraction are among the most challenging phases of SLRs.

3. Methodology

This section outlines the protocol used to perform this tertiary study. The protocol consists of five activities [22], namely defining the research objectives and questions, the search process (terms and resources), inclusion/exclusion criteria, data extraction strategy, and synthesis of the extracted data.

3.1. Research objectives and research questions

To accomplish the goal of this study (see Section 1) we formulate four research questions [3] as listed below:

RQ$_1$: *Does the number of secondary studies explicitly reporting the threats to validity increase over the years, in the software engineering domain?*

By answering this research question, we can find out if there is an increasing awareness of software engineering researchers in reporting the threats to validity of secondary studies. We expect that as the secondary studies community becomes more mature, the frequency of reporting threats to validity is increasing.

RQ$_2$: *What are the most common threats to validity reported by secondary studies?*

RQ$_2$ is related to threats to validity themselves. Specifically, we aim at gathering the most common threats to validity and compile a list of distinct threats to validity. Currently threats to validity are not uniformly reported (i.e., the same threat to validity is reported with a different name by different researchers). Thus, such a list of threats to validity can act as a checklist for authors while designing and conducting secondary studies.

RQ$_3$: *What are the mitigation actions for the most common threats to validity?*

Answering this research question will extend the aforementioned list with the most common ways of mitigating each threat. By browsing this list, researchers will be able to select and apply one or more mitigation actions that will ensure the validity of the planned secondary study. Eventually, this will lead in an increase in the quality of the corpus of secondary studies in the software engineering domain.

RQ$_4$: *What are the most common categorizations (*e.g.*, internal, external, reliability, construct, etc.) of threats to validity for secondary studies?*

RQ$_4$ is related with understanding the nature and types of threats to validity and enhance their reporting. We expect that the comprehensive investigation of threats to validity that will be provided by this study can lead to the development of a schema that can be reused in future secondary studies. Eventually, this is expected to lead to a deeper understanding of the nature of each threat and their effect on the validity of the results.

3.2. Search process

The search process aimed at identifying secondary studies that will be considered as candidates for inclusion in our tertiary study. The procedure consisted of an automated search into well-known digital libraries for publications in specific well-established journals and conferences. The decision to proceed with investigating specific publication venues rather than complete digital libraries means that the coverage of this tertiary study will decrease. However, we preferred to restrict our searching space to well-known journals and conferences so as to obtain a representative sample as suggested by Wohlin et al. [48] and to ensure a higher quality of the /included studies. This is also suggested by Kitchenham et al.: targeted searches at carefully selected venues are justified to omit low quality papers [23]. The proposed research approach, i.e., selecting specific publication venues has been applied in other systematic secondary studies in the field of software engineering (e.g., [6,13,19], etc.), including other tertiary studies (e.g., [20,41,46]).

In addition to selecting only high-quality venues of software engineering research, we have selected to explore only general software engineering venues, and not venues related to software engineering phases (architecture, maintenance, validation and verification, etc.) or application domains (embedded systems, multimedia applications, etc.), so as to reduce bias by the possible maturity of specific communities. Overall, the criteria that were considered while selecting the publication venues are the following:

- We only included venues which are classified "Computer Software" by the Australian Research Council and evaluated higher than or equal to level "B" (for journals) and "A" (for conferences). We consider "Computer Software" because this category includes, among others, the publication venues related to software engineering. Regarding journals, we included "B" because rankings of scientific venues are usually not conclusive and vary between ranking systems. The decision to not include "B" level conferences was taken for two reasons: (a) the number of venues would increase substantially by including "B" class conferences as well and (b) in principal journal publications undergo a more rigorous review process than non-top conferences. Therefore, we opted for the inclusion of only "B" class journals and not conferences.
- Searched venues had to be strictly from the software engineering (SE) domain. The category "Computer Software" also contains venues that do not focus on software engineering. Other venues of very high quality and with a high ranking and a large field rating (such as Communications of the ACM) are excluded since we are only interested in software engineering research; practices from other disciplines might not be applicable in SE.

- We used the Field Rating of venues provided by Microsoft Academic Research[4] as the final criterion for venue quality. More specifically, we exclude venues that do not have a field rating value. Field rating is similar to h-index, since it considers the number of publications and the distribution of citations to them. Field rating only calculates publications and citations within a specific field and shows the impact of the scholar or journal within that specific field. Field rating is to the best of our knowledge the only source where you can extract the same venue quality measures for both journals and conferences.[5] Other measures, such as impact factor or acceptance rates have not been taken into account since they are not uniform across journals and conferences. Furthermore, impact factors and acceptance rates are not available from one common source for all venues but would need to be gathered from different sources, causing threats to the reliability of the study.

The outcome of this process led to the inclusion of the publication venues presented in Appendix C. The results of this selection process, in terms of journals are identical to those of Wong et al. who use the same seven journals for assessing top software engineering scholars and institutions [50]. Concerning conferences, the results are in general in accordance to those of Cai and Card [6], by taking into account that we have excluded conferences specific to development phases. The difference is on the substitution of the Annual Computer Software and Application Conference (COMPSAC) with the International Conference on Software Process (ICSP). COMPSAC is not rated from the Australian Research Council, with an "A" ranking and therefore it was not included in the considered publication venue set. In addition to these publication venues, we have updated our venue selection strategy so as not to only target venues that pass the aforementioned criteria, but also well-established venues that relate to the context of the study (i.e., empirical software engineering). The employed search strategy is already adopted by several secondary studies in software engineering (see [S165]. Thus, we have included Evaluation and Assessment in Software Engineering (EASE) in our searching scope although it failed one criterion (the Field Rating), since we deem it very important in empirical software engineering research. We note that since the focus of the search process is on high-quality studies, all our finding primarily refer and are applicable to high-quality research.

Finally, we only considered the title of the articles, since we aimed at identifying studies that are explicitly aware of the terms literature review and mapping study and categorize themselves as such. Therefore, we queried the digital libraries search engines using the following terms: "survey", "literature review", "mapping study", "mapping studies", "systematic review", "systematic mapping", and "meta-analysis". The term "survey" has been included in the search strategy, since it was the most established unofficial term for literature reviews, before the introduction of the specific terminology. In the secondary literature one can identify search strategies that either target papers' full-text/abstracts [S1], [5], or just the titles of studies [S2], [9]. In the most common case searches that target full-text or abstract are used for narrower research areas that are content-specific, whereas broader topics, similar to our study are more targeted. Additionally, although we acknowledge the fact that some high-quality studies might omit the research method (i.e., literature review or mapping study) from the title of the publication, we believe that this number is rather limited. By manually cross-checking the reference list of a recent tertiary study, we have identified that only 5.5% of studies is missing the research methodology from the title. Furthermore, according to the most common guidelines for performing secondary studies, it is highly recommended to use this terminology in the study's title [18]. Finally, we note that our search string is in complete accordance with a similar tertiary study with a similar objective [5].

As a gold standard for validating our search process we manually cross-checked the reference list of a recent tertiary study (Budgen et al. [5]) and concluded that only 5.5% of studies is missing the research methodology from the title. Additionally we examined the set of secondary studies identified in previous tertiary studies, that were published prior to 2014, in the domain of software engineering [2,9,15,19,20,28] and [41]. In particular, we went through all the secondary studies of the aforementioned tertiary studies, and for those that have been published in the selected venues, we checked if they are part of our secondary study dataset. By following this process, we validated our search process since all papers analyzed in the eight tertiary studies, have been retrieved. We note that this cross-checking included only papers published in journals and conferences that were included in our search process. The article searching process has been performed so as to include all papers published (not accepted for publication) until the end of 2016, i.e., all conferences until the 2016 edition and all journals until December 2016.

3.3. Article filtering phases

The candidate articles that were identified, through the search process described in Section 3.2, underwent an initial exclusion phase, in which we only inspected the abstract. In this phase, all articles that have not been confirmed as Systematic Mapping Studies (SMSs) or Systematic Literature Reviews (SLRs) were excluded. The most common reason for exclusion during this stage was the double meaning of the term "survey" in software engineering bibliography, e.g., "*surveying a population through questionnaires*" [36] instead of "*surveying the literature*" [18].

During the second inclusion/exclusion iteration, we scanned the full-text of the remaining articles and compared them against the following pre-determined criteria:

- Inclusion criteria:
  - ○ Study explicitly discusses threats to validity, in a dedicated paragraph that may appear either in a separate section, or as part of discussion, methodology, etc.
- Exclusion criteria:
  - ○ Study is not a Systematic Mapping Study or Systematic Literature Review. This criterion excluded from the analysis exploratory field studies that have been retrieved through the term "survey" within their title, but refer to the measurement of subjects through questionnaires. Therefore these studies do not include any meta-analysis of primary studies and cannot be considered literature reviews or mapping studies.
  - ○ Study does not describe its own threats to validity, but only of the primary studies.

The set of studies included through this selection process constitute the list of secondary studies investigated in this tertiary study. The list of these references is presented in Appendix A, by providing each study with a unique identifier, used for the rest of the study. A summary of the total and final number of secondary studies retrieved from each venue is provided in Table 3. The article filtering process was performed by the first and the second author independently, and the few disagreements that emerged (namely in 12 studies) were settled rather easily, by considering the inclusion/exclusion criteria. The most common reason for these disagreements was that threats to validity were discussed in a section termed "Limitations", which in some cases refer to "Threats to Validity" and in others to more generic limitations of the study. In the case of a paper reporting two studies (i.e., a secondary and a primary one), the threats to validity section needed to be inspected so as to identify if threats correspond to the secondary or the primary study. An interesting finding from Table 3 is that in only a limited number of publication venues threats to validity tent to be presented in a separate Section.

[4] http://academic.research.microsoft.com/

[5] Google Scholar also provides some related data, but only for 20 venues of the Software Systems category. Therefore, we were not able to extract the data for all candidate venues.

Table 3
Secondary Studies on Software Engineering.

Publication Venue	Initial Search	Final Inclusion
Information and Software Technology	173	73
Journal of Systems and Software	94	30
IEEE Transactions on Software Engineering	41	14
Empirical Software Engineering	63	10
Empirical Software Engineering and Measurement	21	7
Software: Practice and Experience	23	3
International Conference on Software Engineering	16	1
Evaluation and Assessment in Software Engineering	75	27
Other	38	0
Total	540	165

Table 4
Data Analysis Methods.

Research Question	Variables used	Analysis method
RQ1—Frequency of reporting threats to validity per year	[A_2] [A_3]	- Frequency table [A_2] - Line chart [A_2] - Linear Regression [A_2]
RQ2—Most common threats to validity	[A_3]	- Frequency tables for [A_3]
RQ3—Mitigation actions	[A_3], [A_4]	- Crosstabs for [A_3], [A_4]
RQ4—Categorization of threats to validity	[A_3], [A_5]	- Frequency table [A_5] - Crosstabs for [A_3], [A_5]

3.4. Data collection & analysis

On the completion of the study selection phase, we proceeded in building a dataset in order to answer our research questions. During this phase, for each secondary study we collected the following data points:

[A_1] Secondary study title
[A_2] Year published
[A_3] Threats to validity
[A_4] Mitigation actions
[A_5] Explicit categories of threats to validity

The data were independently collected by the first, second and the fourth author. In case of a disagreement, discussions took place until a consensus was reached. The discussion in most of the cases was on the wording used in the retrieved information: in many cases, different wording has been used for the same threat to validity or mitigation action. Therefore, in a large number of cases a disagreement was initially noted, but it was subsequently resolved during the discussion. In the limited number of cases when the disagreement did not stem from textual mismatch, the other two authors were involved so as to resolve the conflict. Data were synthesized using the content analysis method for synthesizing qualitative data [9]. Content analysis is a systematic way of categorizing and coding terms (in our case threats to validity and mitigation actions) by counting and tabulating data. Specifically for our study an iterative process was adopted: when the name of a threat (or mitigation action) was updated, all previous studies that referred to the same threat with the old name were updated to map to the new one. In the end of the data collection process an individual check for synonyms and related threats and mitigation actions that could be further merged was performed by the four senior authors. Similarly as before, the very limited number of conflicts has been resolved in two separate discussion groups among all authors. In Table 4 we provide a mapping of research questions, variables, and the corresponding analysis methods that we used for answering each research question. In this listing [A_1] is not included, as it was not used to answer a specific question, but only for identification reasons. After synthesizing the answers to the individual research questions, we will develop a classification tree that can be considered as the final outcome of this study. The first level of the tree will include the identified categories, the second level will map specific threats to categories, whereas the last level, will list the mitigation actions for a specific threat. Since we acknowledge the subjectivity involved in the aforementioned synthesis process, we: (a) performed a Delphi study with experts on secondary studies and empirical methods, so as to validate the accuracy of our results (see Section 6), and (b) further discuss it as a threat to validity.

4. Results

This section presents the results of this tertiary study, aiming at providing an overview of how threats to validity are identified, categorized and mitigated in secondary studies. The rest of the sub-sections are organized by research question: Section 4.1 presents the frequency with which secondary studies report on threats to validity; Section 4.2, reports the most common threats to validity; Section 4.3 lists the most common mitigation actions for the most threats identified in the previous section; finally, Section 4.4 lists the most common, explicitly mentioned categories for threats to validity, and the specific threats that are mapped to each category. We note that throughout this entire Section 4, the threats are reported exactly in the way that they are presented in the original study, without any synthesis process.

4.1. Awareness on threats to validity (RQ_1)

In order to graphically depict the frequency with which threats to validity are explicitly reported in software engineering secondary studies, we plotted a line chart as presented in Fig. 2. In Fig. 2, the x-axis represents the year, whereas the y-axis the number of published papers. In particular, the dashed line represents the total number of secondary studies that we have identified in our search (i.e., all software engineering secondary studies) whereas the continuous line represents the number of studies that explicitly report threats to validity. The relevant descriptive statistics (i.e., frequency table) are presented in Table 5. We note that although the studies that do not report threats to validity have been excluded from our dataset, we have recorded the number of studies that have been identified per year. We also note that, although the terms Systematic Literature Review (SLR) and Systematic Mapping Study (SMS) were not introduced before 2004, reviews of the literature existed in the research corpus, usually mentioned as “surveys” (see Section 3.1).

From Table 5, we omitted results prior to 2006, because none of the secondary studies published before then, had a dedicated paragraph on

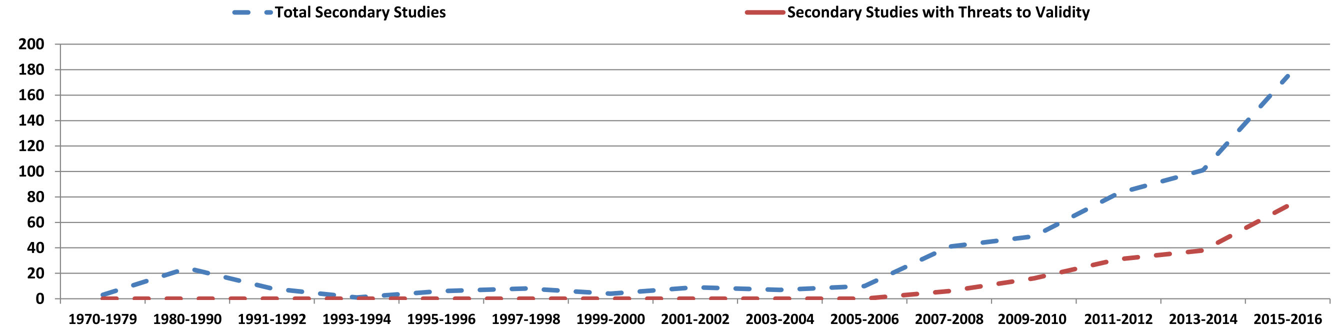

Fig. 2. Studies Reporting Threats to Validity.

Table 5
Secondary Studies in Literature.

Year	Total Studies	Studies with Threats to Validity	Percentage
2007–2008	41	6	14,63%
2009–2010	49	16	32,65%
2011–2012	83	31	37,35%
2013–2014	101	38	37,62%
2015–2016	175	74	41,71%
Total 2007–2016	449	165	36,53%

threats to validity. In Table 5, we present periods of two years, so as to provide a more generic trend without getting influenced by possible outliers. In the last row, we present aggregated values only for the period in which threats have started to be reported (i.e., 2007–2016).

By observing both the Figure and the last column of Table 5, we can recognize an increase in the percentage of secondary studies that report threats to validity. To further explore the frequency of reporting threats to validity, we have tried to identify a trend in the aforementioned data series. For years 2007–2016, with respect to the percentage of studies containing threats to validity with linear regression we have observed the existence of a linear function with slope 7.94% or 0.0794 (p-value < 0.05). Additionally, by performing a One-Sample x^2 test we can observe that the percentage of studies that report threats to validity (from 2003 and on) cannot be captured by a random variability ($p < 0.05$).

The awareness of software engineering researchers on reporting threats to validity for secondary studies is increasing over the years. However, there is still lot of room for improvement, until the community reaches the levels of other, more established empirical research methods.

4.2. Threats to validity (RQ$_2$)

This section aims at presenting the most common threats to validity, as mentioned in the secondary studies. In Table 6 we present threats to validity with a frequency higher than 10 studies.

In order to discuss the threats reported in Table 6, some synthesis activities have been performed. Namely, we merged threats to validity, in a way that they are as specific as possible, while keeping them consistent to the corresponding study. The threats to validity are described as follows (some threats are discussed together):

Table 6
Most Common Threats to Validity.

Threats to validity	Count	Percentage
Study inclusion/exclusion bias	100	17,4%
Construction of the search string	92	16,0%
Data extraction bias	91	15,8%
Selection of DLs	70	12,2%
Researcher bias	40	7,0%
Robustness of initial classification	35	6,1%
Generalizability	27	4,7%
Publication bias	24	4,2%
Repeatability	23	4,0%
Validity of primary studies	13	2,3%
Quality assessment subjectivity	13	2,3%
Coverage of research questions	13	2,3%
Results not applicable to other organizations/domains	12	2,1%
Selection of publication venues	12	2,1%
Search engine inefficiencies	10	1,7%

- **Study inclusion/exclusion bias** (100 studies) refers to problems that might occur in the study filtering phase, i.e., when applying the inclusion/exclusion criteria. Such threats are usually found in studies, in which there are conflicting inclusion/exclusion criteria, or very generic ones.
- **Construction of the search string** (92 studies) refers to problems that might occur when the researchers are building the search string. As a consequence, the search might return a large number of primary studies (including many irrelevant ones) or miss some relevant studies.
- **Data extraction bias** (91 studies) refers to problems that can arise in the data extraction phase. Such problems might be caused from the use of open questions in the collected variables, whose handling is not explicitly discussed in the protocol. A special type of data extraction bias is the **Quality assessment subjectivity** (13 studies), i.e., the process during which the quality of the primary studies is evaluated by the authors of the secondary study. This threat is relevant only for SLRs that report the evaluation of primary studies' quality.
- **Selection of Digital Libraries (DL)** (70 studies) refers to problems that can arise from using very specific, too broad, or not credible search engines. The consequence of this threat can be either the return of a lot irrelevant or the miss of relevant studies. In addition to that **Search Engine Inefficiencies** (10 studies) pointed out cases when the search engine interface cannot accommodate complex queries.
- **Researcher bias** (40 studies) refers to potential bias the authors of the secondary studies may have, while interpreting or synthesizing the extracted results. This can be a bias towards a certain topic, or because only one author worked on data synthesis.
- **Publication bias** (24 studies) refers to cases where the majority of primary studies are identified in a specific publication venue. For example, if the majority of primary studies stem from a single workshop, the likelihood of biasing the results, based on the beliefs of a certain community, is rather high. Another type of publication bias is the **Validity of the primary studies** (13 studies), which suggest that the results of the secondary study might be biased from inaccurate

results reported in the primary studies. A common reason for this is that studies with negative results are less probable to get accepted for publication.

- **Robustness of initial classification** (35 studies) is applicable to secondary studies, whose data collection relies upon a classification schema. A common practice while performing such an activity is to identify an existing classification schema that (if needed) is tailored to fit the needs of the secondary study. The selection of this initial classification schema poses a threat to validity, since it might not be fitting for the domain, and its tailoring is not efficient.
- **Generalizability** (27 studies) refers to the possibility of not being able to generalize the results of the secondary study (for example due to the identification of only a portion of existing primary studies). A special case of this threat that is quite frequently reported is **Results not applicable to other organizations or domains** (12 studies).
- **Repeatability** (20 studies) refers to threats that deal with the replication of a secondary study. The most common reason for the existence of such threats is the lack of a detailed protocol, or the existence of researcher and data extraction bias.
- **Coverage of Research Questions** (13 studies) refers to the set of research questions not adequately fulfilling the goal of the secondary study. Possible reasons are setting a very generic goal, or the improper decomposition of the goal into questions.
- **Selection of publication venues** (12 studies) refers to the problem that might occur, when the research team selects to explore specific venues rather than using broad search engines. The most common rationale for this decision is either the fact that a topic is too broad, or if the research aims at high quality studies only. The consequence of this threat is the miss of relevant studies.

By analyzing the aforementioned dataset from the perspective of the total count of reported threats to validity, we have observed that in secondary studies, on average, 4.36 threats to validity are reported. The outcomes show that the *minimum* number of threats recorded is 1 (since studies without threats have been excluded from our analysis), the *maximum* number is 9, the *median* value is 4 threats, the *mode* value is 3 threats and the *std. deviation* is 1.59. We have identified only one study as an outlier (reporting 9 threats to validity), but its influence on the average value is very limited and therefore we have not removed the study from the analysis.

4.3. Mitigation actions (RQ_3)

In this section we report the most common mitigation actions for the most common threats to validity, namely threats that have been reported in more than 15 studies (see Table 6). We note that in some cases the same mitigation action is connected to more than one threat. The mapping of mitigation actions to threats to validity is presented in Table 7. In particular for every threat to validity we present a list of mitigation actions, and the number of studies in which they are applied (in parenthesis). The full list of mitigation actions (more than 500 distinct actions) for all threats to validity has been omitted from this manuscript, due to page limitations, but is available in the accompanying technical report.[6]

Due to space limitations the discussion of all mitigation actions is not possible. Thus, for every threat presented in Table 7, we discuss the top-3 most frequently occurring mitigation actions:

- ***Study inclusion/exclusion bias*** is mitigated by discussion among the authors and by employing an external opinion for resolving disagreements. In addition, the inclusion and exclusion criteria should be clearly defined in a protocol, which is updated along the whole process.
- In order to mitigate threats related to ***construction of the search string*** usually snowballing is executed. Snowballing is a technique that is attempting to identify missed studies, based on the reference list of already obtained papers. A detailed guidance on how to apply the snowballing technique has been provided by Wohlin [49]. As an alternative, authors consider synonyms and continuously refine their search process.
- ***Data extraction bias*** is mitigated by discussing the data during the recording process, or by introducing a cross-check of the extracted data from a more senior researcher. The cross-check process implies that a senior researcher that was not involved in the original data extraction, validates the initially extracted data. This cross-check should be performed on a portion of the dataset. During this process the role of the additional researchers is to cross-check results, or resolve conflicts.
- ***Selection of digital libraries*** is mitigated through the inclusion of the most well-known digital libraries. This process involves the selection of venues or digital libraries that are the most established in the field of research. According to Kitchenham et al. [23], both generic-scope and domain-specific venues should be considered. The most commonly used databases are: ACM, IEEE, ScienceDirect, Springer, Scopus, Web of Science, and Wiley [23]. In case the research team is investigating a very broad topic, or is interested in including only top quality venues, venue selection processes are described in [6,13], and [19]. To avoid this threat, some authors select to explore specific venues, or others to use broad search engines and indexes (e.g., Google Scholar, Scopus, etc.)
- To mitigate ***Researchers' bias*** secondary studies' authors discuss the interpretation of the results, and perform pilot data analysis. Also reliability analysis and cross checks have to be performed.
- Concerning ***publication bias***, authors use snowballing, scan selected venues, or include grey literature. Also, expert opinion can be used to assess the extent to which the study is subject to publication bias.
- Regarding the ***robustness of initial classification***, existing studies suggest its extensive discussion/cross-checking between the researchers, or the use of existing/well-defined classification schemas.
- To mitigate ***lack of generalizability***, use of broad searches and comparison to results of other studies.
- Secondary studies ***repeatability*** is assured with the development of a protocol that reports the use of a systematic process that can be followed, or the clear definition of search terms and procedures. The process should follow well-defined guidelines.

The average number of mitigation actions per study is 6.27, the *minimum* number is 0, the *maximum* number is 17, the *median* and the *mode* value is 6 mitigation actions and the *std. deviation* is 2.9. The average number of mitigation actions per identified threat to validity is 1.54, the *median* and the *mode* value is 1 mitigation action per thread and the *std. deviation* is 1.13. From the aforementioned results we can observe that: (a) some actions (e.g., *Inclusion of most known digital libraries* and *manual search of publication venues*) can be used to mitigate two threats—*inadequacy of initial publications identification* and *lack of generalizability*; (b) the threats to validity that were least often mitigated are *publication bias*, and *generalizability*; and (c) the mitigation actions *discuss* and *cross-check* are very generic and fit almost every threat to validity, e.g., discuss the extracted data, or cross-check the data selection.

Another interesting observation that stems from the answer to this research question is the cost of applying a mitigation action. For example a mitigation action that can be performed early in the review process (e.g., setting concrete inclusion/exclusion criteria) is less expensive (in terms of effort) than discussions among authors in data extraction. To this end, we propose that researchers prioritize mitigation actions that are applicable to early review phases, rather than postponing validity assessment for later stages. In any case, according to various guidelines for empirical software engineering validity management is part of the

[6] http://se.uom.gr/wp-content/uploads/IST_material.zip

Table 7
Most Common Mitigation Actions.

Threat to Validity: Study inclusion/exclusion bias Discussion of marginal cases (44, 6.9%) Definition of inclusion / exclusion criteria in a protocol (22, 3.5%) Revision of inclusion / exclusion criteria (16, 2.5%) Employment of a third opinion for marginal cases (10, 1.6%) Employment of a systematic voting approach (8, 1.3%) Cross-checking of paper selection (6, 0.9%) No mitigation (8, 1.3%)[a] Use of random paper screening (5, 0.8%) Execution of a consensus meetings (2, 0.3%) 32 other actions encountered once (5%)	**Threat to Validity**: Construction of the search string Employment of snowballing (27, 4.3%) Inclusion of synonyms/roots (20, 3.2%) Use of a gold standard (11, 1.7%) Systematic search string construction (13, 2.1%) Constant search string refinement (10, 1.6%) Extension of search scope / Broad terms (10, 1.6%) Execution of pilot searches (9, 1.4%) No mitigation (9, 1.4%) Use from previous studies (4, 0.6%) Use of author and citation analysis (3, 0.5%)
Threat to Validity: Data extraction bias Discussion among authors (30, 4.7%) Involvement of more researchers / Work in pairs (15, 2.4%) Use of a data extraction form (12, 1.9%) Cross-checking of data extraction (11, 1.7%) Use of random paper screening (11, 1.7%) Execution of pilot data extraction (10, 1.6%) No mitigation (9, 1.4%) Employment of a third opinion for conflicting data items (8, 1.3%) Definition of a review protocol (5, 0.8%) Use of Codes (3, 0.5%) Ensure the conformance to guidelines (3, 0.5%)	**Threat to Validity**: Selection of DLs Inclusion of most known DLs (35, 5.5%) Use search engines and indexes (14, 2.2%) Employment of snowballing (13, 2.1%) Inclusion of specific venues (10, 1.6%) No mitigation (7, 1.1%) Use of expert opinion (2, 0.3%) Consideration of a large time period (2, 0.3%) Inclusion of grey literature (2, 0.3%) Ensure the conformance to guidelines (2, 0.3%)
Threat to Validity: Robustness of initial classification Use an existing classification scheme (15, 2.4%) Discussion among authors (7, 1.1%) Employment of a third opinion for the classification (5, 0.8%) No mitigation (4, 0.6%) Application of keywording of abstracts (4, 0.6%)	**Threat to Validity**: Researcher bias Discussion among authors (16, 2.5%) No mitigation (13, 2.1%) Execution of pilot data analysis (6, 0.9%) Use of reliability checks (4, 0.6%) Development protocol (3, 0.5%) Comparison with existing studies (3, 0.5%)
Threat to Validity: Repeatability Development of a review protocol (8, 1.3%) Ensure the conformance to well-established guidelines (7, 1.1%) Documentation of the search process (5, 0.8%) Involvement of more than one researcher in the process (3, 0.5% Documentation of inclusion/exclusion criteria (2, 0.3%) Documentation of the review process (3, 0.5%) Ensure the public availability of data (2, 0.3%) No mitigation (2, 0.3%)	**Threat to Validity**: Publication bias No mitigation (14, 2.2%) Inclusion of grey literature (5, 0.8%) Use of broad time and publication coverage (3, 0.5%) Scanning of selected venues (2, 0.3%) Use of expert opinion (2, 0.3%) 5 other actions encountered once (0.8%)
Threat to Validity: Generalizability No mitigation (14, 2.2%) Use of broad time and publication coverage (4, 0.6%) Comparison to other studies (3, 0.5%) Use both academic and industrial papers (2, 0.3%)	

[a] This refers to cases when a threat to validity is reported in a secondary study, but no mitigation action is referenced to resolve it.

empirical study protocol and should be assessed before conducting the study.

4.4. Threats to validity categories (RQ$_4$)

In Table 8, we report the most commonly used categories for classifying threats to validity (as reported by the authors of the secondary studies). In this table, we have omitted categories of validity that are found in only one study. From the results, we can observe that the majority of the reported threats, i.e., 61.4%, are not classified into any category; whereas, 28.8% of the studies reported the corresponding threats to validity based on the guidelines of Wohlin et al. (i.e., conclusion, internal, construct, and external validity) [47]. Furthermore, we observe the existence of categories that are specific for secondary studies (i.e., *data extraction, primary studies identification*, and *publication bias*); such categories have not been used in the past to report threats to validity in empirical software engineering (see Section 2.1). We note that *objectivity* appears to be a superset of *data extraction, data interpretation*, and any other activity that may introduce bias. We emphasize again that these categories are listed here as reported in the secondary studies, even though they could possibly be classified into the categories of Table 1 (for example *generalization* is similar to *external*, while *data extraction* belongs to *reliability*). Finally, all of these categories have already been reported as specific threats to validity by other studies (see overlap with Table 6); this suggests that the threshold of granularity between a single threat and a category of threats is not clear.

Table 8
Explicit Categories of Threats to Validity.

Explicit Categories	Count
Not defined	329
Construct	54
Internal	51
External	37
Reliability	24
Conclusion	23
Primary study identification	9
Generalization	7
Data extraction	7
Theoretical	5
Objectivity	5
Publication bias	5
Interpretive Validity	3

Table 9
Classification of Threats to Categories.

	Categories									
Threats	Reliability	Primary study Identification	Objectivity	Data Extraction	Internal	Generalization	External	Construct	Conclusion / Interpretive	% of Dominant Category
Generalizability	1	0	0	0	0	1	20	1	0	87,0%
Search string	0	7	0	2	5	0	1	23	0	60,5%
Selection of DLs	0	1	0	0	4	0	1	10	0	62,5%
Publication bias	0	0	0	0	2	0	1	1	0	50,0%
Publication venues	0	0	0	0	1	0	0	2	0	66,7%
Data extraction	2	0	1	2	12	1	1	2	4	48,0%
Researcher bias	1	0	0	2	3	1	0	2	6	40,0%
Repeatability	11	0	0	1	0	0	1	0	5	61,1%
Quality assessment	0	2	1	0	2	0	0	0	1	33,3%
Research questions	0	0	0	0	0	0	1	5	0	83,3%
Primary studies	2	0	0	1	3	0	0	0	0	50,0%
Study selection	4	6	2	0	13	0	0	6	2	39,4%
Initial classification	1	0	1	1	3	0	0	7	2	46,7%

Another observation that further unveils the confusion in the community, while reporting threats to the validity of secondary studies, is the overlap that exists in the classification of the most common threats. In Table IX, we present the cross-tabulation of threats to validity and their categories, again as reported by the authors of the corresponding secondary studies. The dominant category where each threat is classified is highlighted in bold font.

From the results of Table 9, we can observe that there is no threat to validity that is always classified under one category by all researchers. For example (a rather uniform case), the *construction of the search string* is in 86% of the cases characterized as a threat to *construct* validity; however there are other studies that classify it as either *internal* threat or *primary study identification* threat. On the other hand (a rather conflicting case), the *study selection bias* is classified as an threat to *internal* validity in 38% of the studies, and as *reliability, primary study identification, objectivity, construct,* or *conclusion* validity threat by the rest of the studies. On average, 59% of the cases are classified in the dominant category (see last column of Table 9).

5. Discussion

The identification, categorization and mitigation of threats to validity is an important part for secondary studies. During the last decade, the ratio of secondary studies managing threats to validity has continuously increased. However, our results suggest that a considerable confusion still exists in terms of terminology, mitigation strategies, and classification. We further focus on the ***classification*** of threats to validity and consider the example of the *study selection bias* threat, which is classified under *internal* validity almost as often as under *reliability*. Arguably, problems in study selection can threaten both aspects of validity. On the one hand, if some studies are falsely included / excluded, the examined dataset will not be accurate (internal validity). Therefore, the investigation of any relationship will be prone to erroneous results. On the other hand, failing to include some studies in the final selection can greatly reduce the possibility that an independent replication reaches the same results (reliability). While one can argue about the correctness of both classifications, having more than one classification can be confusing and does not allow for a uniform comparison of the threats. We therefore argue that a new uniform classification schema is required.

The rest of the section is organized as follows: In Section 5.1, we present and discuss the proposed classification schema for threats to validity of secondary studies. In order to facilitate the usability of the classification schema, in Section 5.2 we compile a checklist that can be used by authors of secondary studies, so as to assess the validity of their study. We note that the threats to validity and mitigation actions reported in this Section are produced as a result of a synthesis process and therefore slightly deviate from those presented in Section 4.

5.1. Classification schema

Our aim is to construct a classification schema for threats to validity that is tailored for secondary studies. According to Nickerson et al., the most common method for building classification schemas for information systems is the three-level indicators model, which is based on both empirical and deductive approaches [30]. We apply this model by: (a) examining the objects (i.e. studies), (b) identifying general distinguishing characteristics of the objects (see results presented in Section 4), and (c) grouping their characteristics so as to create our classification schema. Specifically, in step (b) we identified three characteristics that will constitute the three levels of the proposed schema: the first one depicting threat categories, the second, threats per se, whereas the third one, mitigation actions.

In order to derive the threat categories (first level of the schema) we used the planning phases of the secondary studies (i.e., search process, study filtering, data extraction, and data analysis – see Fig. 3), instead of using the aspects of validity that are threatened (e.g., internal/external/construct validity, etc.). In addition to this, we have added an additional category (i.e., a horizontal one) that corresponds to threats that cover the complete lifecycle of the secondary study. Thus, the threat categories for our schema are the following:

- ***Study Selection Validity***. This category involves threats that can be identified in the first two phases of secondary studies planning (i.e., search process and study filtering phase). Issues classified in this category threaten the validity of searching and including primary stud-

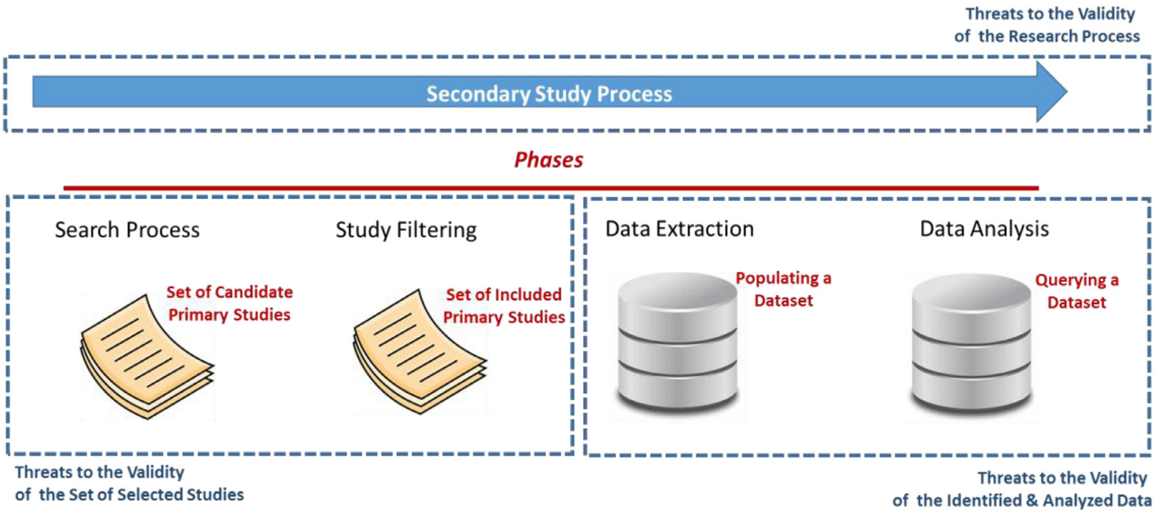

Fig. 3. Secondary Studies Phases and Corresponding Threats.

ies in the examined set. This involves threats like the *selection of digital libraries, search string construction*, and *study selection bias*, etc.

- ***Data Validity***. This category includes threats that can be identified in the last two phases of secondary studies (i.e., data extraction and analysis) and threaten the validity of the extracted dataset and its analysis. Examples of threats in this category are *data collection bias, publication bias*, etc.
- ***Research Validity***. Threats that can be identified in all four phases and concern the overall research design are classified into this category. Examples of threats falling in this category are: *generalizability, coverage of research questions*, etc.

Although we believe that the current classification schema improves the orthogonality among threat categories: (a) there are still some grey-zone threats (see bullet list in page 20), (b) there are some cause-effect relationships between threats. First, using the proposed classification schema, we address the problem of classifying a single threat to two categories (e.g., as mentioned at the beginning of Section 5): every threat is classified within one category, based on the phase of the study design where it was identified and the set of artifacts whose validity is threatened. We identified only five cases that seem to be on a “grey zone” between two categories:

- ***Quality Assessment Subjectivity***—Quality Assessment in some cases (based on the secondary study design) can act as a means for study selection (i.e., in cases when a specific level of quality needs to be assured for included primary studies); in others it acts as part of data extraction (i.e., in cases when the assessment of quality of the primary studies is part of the research questions of the study). Thus, *Quality Assessment Subjectivity* can be classified in both Study Selection Validity and Data Validity, based on the role of the quality assessment. Thus, for SLRs, this threat is normally classified as a threat to Study Selection Validity, whereas for Systematic Mapping Studies, it can be classified as a threat to Data Validity. To ease the readability of this section, *Quality Assessment Subjectivity* will be presented only as part of *Data Validity*.
- ***Publication Bias*** and ***Validity of Primary Studies*** —Although *Publication Bias* and *Validity of Primary Studies* stem from the study selection phase, they threaten the validity of the extracted data, their analysis, and the subsequent interpretation. In particular *publication bias* may result in an extracted dataset that does not represent a wide research community, but only reflects the opinions of a limited number of researchers. Furthermore, low validity of primary studies threatens the validity of the extracted dataset, since they may offer low-quality evidence. Thus, we have classified both threats in the *Data Validity* category.
- ***Robustness of initial classification*** and ***Construction of attribute framework***. These two threats are highly related to data validity in the sense that if a ‘wrong’ classification schema is selected the complete data collection will be misguided due to the use of inaccurate classification classes and terminology. Thus, the correctness of the final dataset is threatened. Although these threats first appear in the study selection phase their impact is mainly observed in the Data analysis phase.

Additionally, one can suggest that a cause-effect relation exists between some threats to validity. For example, if a search process is based upon specific search terms and some are overlooked (study selection validity), the results may not be generalizable in a wider population (research validity). Thus, the first two categories (study selection and data validity) correspond to the phase when a threat is introduced (e.g., search string construction), whereas some research validity threats concern the actual impact of that threat. In such cause-effect relations the impact is mostly on generalizing the results. For example, by introducing an error in the search process (study selection validity) we cannot generalize to the population of the studies (research validity).

Next, each category of threats is discussed in detail, based on the findings reported in Section 4 (i.e., purely based on the extracted data). We note that due to space limitations, only the most frequent mitigation actions for every threat are presented in Fig. 4a–c. The full list of mitigation actions is available online, in the accompanying technical report. The three categories of validity threats along with the proposed mitigation actions are presented in Fig. 4a–c. The light blue rounded rectangles represent threats to validity, whereas pink rounded rectangles correspond to mitigation actions. Dotted lines are used to depict threats that can be grouped together under a more generic threat. Also dotted lines are used to group together mitigation actions that all are used to minimize a possible threat.

The ***study selection validity*** category involves 11 specific threats (see Fig. 4a). Five threats to validity (see top part of Fig. 4a) can be grouped in a more generic one, i.e., *Adequacy of initial relevant publication identification*, whereas the rest are ungrouped. From the threats of this category, some are mutually exclusive, whereas others are complementary. For example, if *selection of digital libraries* is performed, the threat *selection of publication venues* is excluded since, normally only one of the two search strategies is selected (except if a quasi-gold standard from specific venues is used for study selection validation; then both strategies are used). The *construction of the search string* threat exists both when DLs or specific publication venues are selected. After the initial set of publications is derived, other aspects threaten the validity of the study: *how have the authors handled the duplicate articles* or *the grey literature, what*

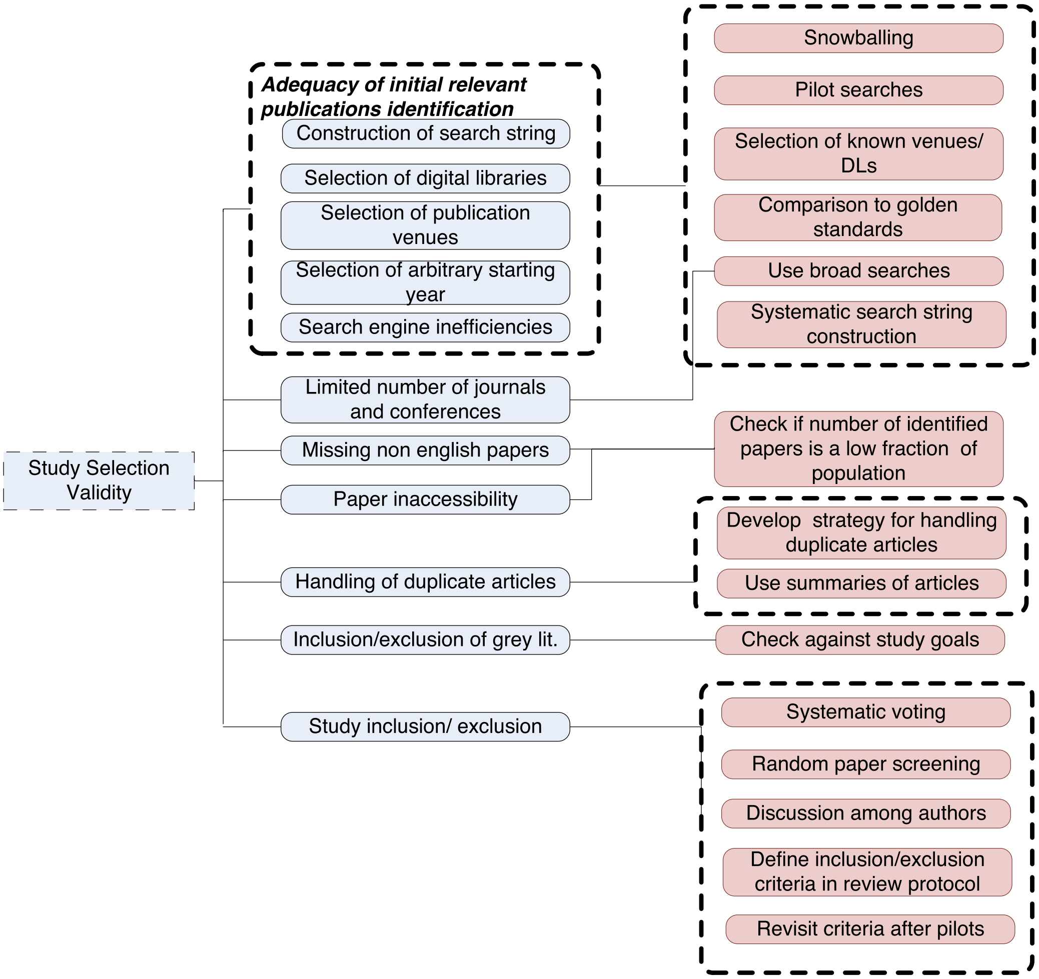
Study Selection Validity
Adequacy of initial relevant publications identification
Construction of search string
Selection of digital libraries
Selection of publication venues
Selection of arbitrary starting year
Search engine inefficiencies
Limited number of journals and conferences
Missing non english papers
Paper inaccessibility
Handling of duplicate articles
Inclusion/exclusion of grey lit.
Study inclusion/ exclusion
Snowballing
Pilot searches
Selection of known venues/ DLs
Comparison to golden standards
Use broad searches
Systematic search string construction
Check if number of identified papers is a low fraction of population
Develop strategy for handling duplicate articles
Use summaries of articles
Check against study goals
Systematic voting
Random paper screening
Discussion among authors
Define inclusion/exclusion criteria in review protocol
Revisit criteria after pilots

Fig. 4a. Study Selection Validity Threats.

languages have the authors explored, were all papers accessible by the authors, were there enough journals and conferences for the authors to search, and is the *selection of inclusion/exclusion criteria accurate*? Threats that appear in Fig. 4a and have not been discussed in Section 4.2, are outlined in Appendix A.

The ***data validity*** category includes 15 specific threats (see Fig. 4b), that are organized into three groups and five ungrouped threats to validity. The first group (middle part of Fig. 4b) includes any kind of bias that can be introduced while collecting data, namely: *data extraction bias, data extraction inaccuracies, quality assessment subjectivity, unverified data extraction*, and *misclassification of primary studies* (mostly relevant for mapping studies). The second group (see top part of Fig. 4b) includes limitations of the dataset that are due to the nature of the subject and not due to researchers' bias (i.e., *small sample size* and *heterogeneous primary studies*). The third group (see bottom part of Fig. 3b) represents threats that are relevant for mapping studies and have been posed by the use of *inadequate classification schemas* or *attributes frameworks*. Furthermore, other aspects such as the *validity of primary studies, the potential lack of relationships in the dataset, the publication bias*, and *the choice of extracted variables* are classified in this category since they are prone to damaging the quality of the dataset. Other individual threats that are mapped to this category are: the researchers' bias while interpreting the results and the lack of statistical analysis. The threats to validity that appear in Fig. 3b and have not been discussed in Section 4.2 are outlined in Appendix A).

Finally, the ***research validity*** category includes 8 specific threats (see Fig. 4c) that are forming two groups and include four ungrouped threats. The first group (see top part of Fig. 4c) represents threats that have to do with the followed process. First, there is a possibility that the *selected research method* (i.e., mapping study vs. literature review) does not fit the goal of the study. Second, sometimes researchers *deviate from the established review process*. The second group (see bottom of Fig. 4c) involves threats to *generalizability*. The individual threats that are mapped to this category are the *lack of comparable studies*, the *coverage of research questions*, and the *unfamiliarity of researchers with the application domain*. Finally, *repeatability* has been classified in this category since although it is threatened by data unavailability; it is also threatened by any undocumented parts of the reviewing process. Therefore, it is considered more as a horizontal threat (that pertains to the whole research process), rather than a specific threat for the data extraction or analysis phase. The threats to validity that appear in Fig. 4c and have not been discussed in Section 4.2 are outlined in Appendix A.

Concerning the mapping of mitigation actions to specific threats to validity, our classification has revealed an interesting relationship. The threats that are grouped together can be mitigated with similar actions. For example, *Snowballing* is used as a mitigation action for three threats to validity of the *Study Selection* category: *construction of the search string, selection of DLs*, and *selection of inclusion/exclusion criteria.* In addition, we can observe that some mitigation actions (e.g., *develop a protocol*) are more generic, in the sense that they alleviate a number of different validity threats (e.g., mitigating the majority of Research Validity threats).

By comparing the findings reported in Fig. 4 to the quality assessment criteria derived from the medical science and the guidelines for

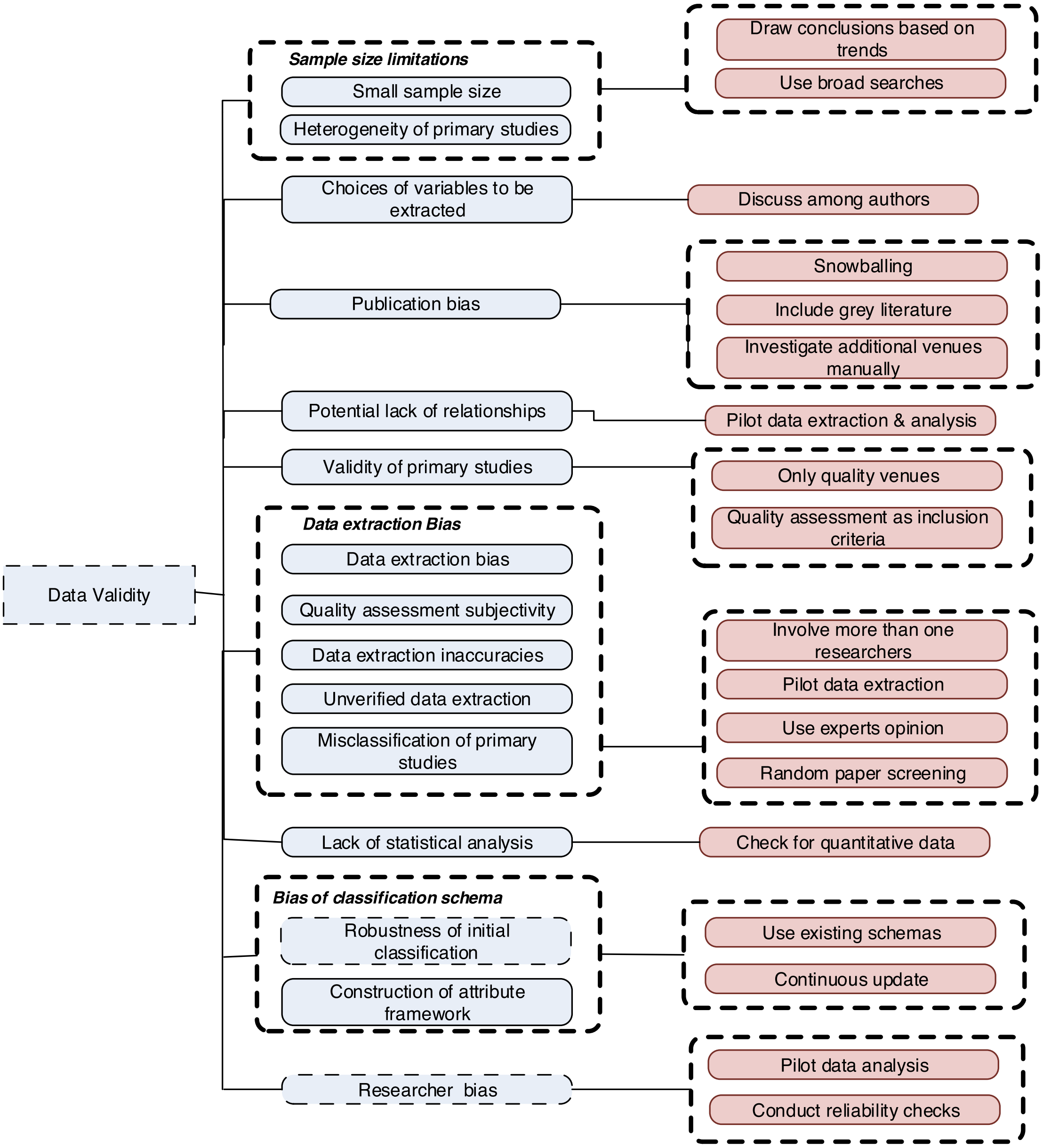
Data Validity
Sample size limitations
Small sample size
Heterogeneity of primary studies
Draw conclusions based on trends
Use broad searches
Choices of variables to be extracted
Discuss among authors
Publication bias
Snowballing
Include grey literature
Investigate additional venues manually
Potential lack of relationships
Pilot data extraction & analysis
Validity of primary studies
Only quality venues
Quality assessment as inclusion criteria
Data extraction Bias
Data extraction bias
Quality assessment subjectivity
Data extraction inaccuracies
Unverified data extraction
Misclassification of primary studies
Involve more than one researchers
Pilot data extraction
Use experts opinion
Random paper screening
Lack of statistical analysis
Check for quantitative data
Bias of classification schema
Robustness of initial classification
Construction of attribute framework
Use existing schemas
Continuous update
Researcher bias
Pilot data analysis
Conduct reliability checks

Fig. 4b. Data Validity Threats.

conducting secondary studies in software engineering, we have identified that some best practices are currently not (or at least not frequently) applied. Nevertheless, we need to note that the majority of mitigation actions reported in the software engineering guidelines and the medical quality assessment instruments are being followed already (e.g., snowballing, handling of duplicate papers, involvement of more than one researchers in data extraction, development of a protocol, etc.). The overlooked best practices are summarized below:

Study Selection Validity

- *Adequacy of initial relevant publications identification*: The search process should be ***reviewed by independent experts*** [24,34] before it is conducted. After the retrieval of the candidate primary studies dataset, it is highly advised to ***evaluate the search results*** [34]. An example of such a process is the gold standard comparison, which is included in Fig. 4a. Nevertheless, we need to note that additional ways to check the fulfillment of this objective can be used. Furthermore, the search process can become more sophisticated by using dedicated ***tools for bibliography management*** [18] (e.g., JabRef, Zotero, etc.) and by ***continuously documenting the search process*** (all stages) [5,18], designating which papers are being excluded and based on which exclusion criterion.
- *Study inclusion / exclusion bias*: In addition to all mitigation actions reported in Fig. 4a, a more formal inclusion / exclusion process can be supported by using pre-defined set of ***decision rules*** [34]. A subset of such rules could dictate how conflicts are being resolved, what is the tolerance in level of disagreement, etc. In this context the most common measure for capturing disagreement is the assessment of the ***kappa statistic*** [18], which is used in the large majority of secondary studies. In addition, in cases when primary study quality is an inclusion/exclusion criterion, secondary studies/ guidelines suggest the definition of a clear ***threshold for study inclusion*** [18,41]. Furthermore, it highly advisable that before executing the study inclusion/exclusion process, an ***independent researcher reviews*** the corresponding part of the protocol (i.e., inclusion/exclusion process)

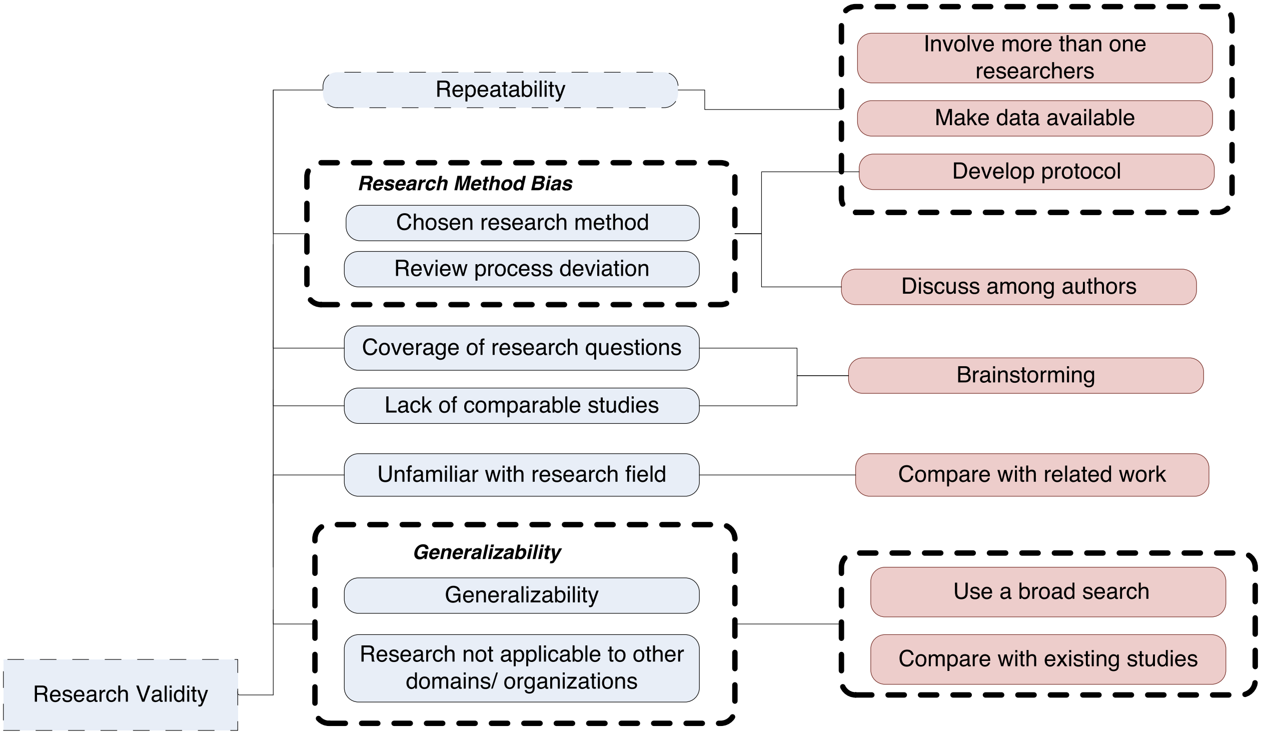
Repeatability
Research Method Bias
Chosen research method
Review process deviation
Coverage of research questions
Lack of comparable studies
Unfamiliar with research field
Generalizability
Generalizability
Research not applicable to other domains/ organizations
Research Validity
Involve more than one researchers
Make data available
Develop protocol
Discuss among authors
Brainstorming
Compare with related work
Use a broad search
Compare with existing studies

Fig. 4c. Research Validity Threats.

[24,34]. Finally, to measure the variability of the results caused by missing studies, ***sensitivity analysis*** can be performed [18].

Data Validity

- *Validity of Primary Studies*: Based on the medical quality assessment instrument, the validity of the primary studies included in the analysis, along with their impact, should be assessed with the ***use of statistical methods***, [1,10,30,45] i.e., cumulative meta-analysis, funnel plots, etc.
- *Data Extraction Bias*: Similarly to inclusion /exclusion bias, the use of the ***kappa-statistic*** [18,24] is highlighted as important for identifying cases, in which researchers' opinions differ. In addition to that, in the special case of performing a mapping study it is advisable to use ***keywording of abstract*** [33] as a means for more efficient data extraction.
- *Researcher Bias*: Regarding researcher bias introduced during the data synthesis stage, it is recommended to adopt ***formal research synthesis methods***[7] (e.g., grounded theory, meta-ethnography, narrative synthesis, etc.). Additionally, the medical research guidelines suggest ***using the scientific quality of primary studies*** [39,43] appropriately, while formulating conclusions. Finally, ***sensitivity analysis*** [18] can be used for measuring the impact of researchers' bias in the extracted conclusions.

Research Validity

- *Repeatability*: To enhance repeatability it is important to ***precisely report the complete process of the review*** [5,18,41], not focusing only on the protocol, but also documenting aspects of the "conducting the review" phase. In particular, it is important to document all the attributes reported in Fig. 1, representing the "reporting" phase.
- *Coverage of Research Questions*: The correct identification of research questions is of paramount importance for a successful secondary study. In particular, it is suggested to ***motivate the need and relevance of the review, as well as each research question independently*** [18,24,34]. To achieve this, apart from having a deep knowledge of the corresponding literature, it is advised to ***consult the target audience*** [34].

The aforementioned best practices, along with the outcomes of the tertiary study (see Fig. 4) are compiled in a checklist which is the final outcome of this work, presented in Section 5.2.

5.2. Checklist for threats to validity identification and mitigation

In this section, and based on the classification schema of Fig. 3, we present a checklist (as a series of questions) that authors of secondary studies should answer when performing secondary studies, so as to assess the validity of their studies. This instrument can aid both in the identification of threats (since not all threats apply in all studies) and the suggestion of mitigation actions (what the authors can do if they identify any threat in their study design). We note that this checklist does not provide additional information compared to the classification schema of Section 5.1, but only acts as a different view of the obtained results. We offer this checklist as a more usable view that can be directly exploited by authors of secondary studies.

The structure of the checklist is quite simple: First a question is asked to understand if a specific threat exists (TV_n), and then a series of sub-questions are asked to check if a proper mitigation action MA_m has been performed. The numbering of mitigation actions is restarted for every threat to validity. Each of the three boxes below corresponds to one category of threats: study selection, data and research validity. For example, $TV_1 - TV_7$ correspond to the seven threats that are reported in Fig. 4a (study selection validity). The mapping between questions and threats reported in Fig. 4 is one-to-one, by considering the groups discussed in Section 5.1. In addition, in a parenthesis following each mitigation action we denote if the action is preventive (P) or corrective (C), i.e., if the action prevents the occurrence of the threat, or corrects / evaluates its importance after its identification.

[7] Research synthesis is "*a collective term for a family of methods that are used to summarize, integrate, combine, and compare the findings of different studies on a specific topic*" [9].

Study Selection Validity

TV$_1$: Has your search process adequately identified all relevant primary studies?

MA$_1$: Have you used snowballing? (P)

MA$_2$: Have you performed pilot searches to train your search string? (P)

MA$_3$: Have you selected the most-known DLs *or* have you made a selection of specific publication venues *or* used broad search engines or indices (*based on the goal of your study*)? (P)

MA$_4$: Have you compared your list of primary studies to a gold standard or to other secondary studies? (C)

MA$_5$: Have you used a broad search process in generic search engines or indices (e.g., Google Scholar) so that you ensure the identification of all relevant publication venues? (P)

MA$_6$: Have you used a specific strategy for systematic search string construction? (P)

MA$_7$: Has an independent expert reviewed the search process? (P)

MA$_8$: Have you used tools to facilitate the search process? (P)

MA$_9$: Have you evaluated the search results and documented the search outcomes? (P)

TV$_2$: Were primary studies relevant to the topic of the review published in several different journals and conferences?

MA$_1$: Have you used a broad search process in generic search engines or indices (e.g., Google Scholar) so that you ensure the identification of all relevant publication venues? (P)

TV$_3$: Have you identified primary studies in multiple languages?

MA$_1$: Is the number of such studies expected to be high compared to the population? (C)

TV$_4$: Were the full texts of all identified primary studies accessible from the researchers?

MA$_1$: Is the number of studies with missing full texts expected to be high compared to the population? (C)

TV$_5$: Have you managed duplicate articles?

MA$_1$: Have you developed a consistent strategy (e.g., keep the newer one *or* keep the journal version) for selecting which study should be retained in the list of primary studies? (P)

MA$_2$: Have you used summaries of candidate primary studies to guarantee the correct identification of all duplicate articles? (P)

TV$_6$: Have you included/excluded grey literature?

MA$_1$: Does your decision to include or exclude the grey literature comply with the goals of the study *and* the availability of sources on the subject? (C)

TV$_7$: Have you adequately performed study inclusion/exclusion?

MA$_1$: Have you used systematic voting? (P)

MA$_2$: Have you performed a random screening of articles among all authors? (P)

MA$_3$: Have researchers discussed the inclusion or exclusion of selected articles in case of conflict? (P)

MA$_4$: Have the inclusion exclusion criteria been documented explicitly in the protocol? (P)

MA$_5$: Have the authors discussed the inclusion/exclusion criteria *and* revised them after pilot iterations, or by experts' suggestions after review? (P)

MA$_6$: Have you prescribed a set of decision rules for study inclusion/exclusion? (P)

MA$_7$: Have you defined quality thresholds for inclusion/exclusion? (P)

MA$_8$: Have you performed sensitivity analysis? (P)

MA$_9$: Have you identified experts' disagreement level with the kappa statistic? (P)

Data Validity

TV$_8$: Is your sample size large enough so that the obtained results can be considered valid?

MA$_1$: Have you tried to draw conclusions based on trends? (C)

MA$_2$: Have you used a broad search process in generic search engines or indices (e.g., Google Scholar) so that you ensure the identification of all relevant publication venues? (P)

TV$_9$: Have you chosen the correct variables to extract?

MA$_1$: Has the choice of variables been discussed among authors, so as to guarantee that the set of research questions can be answered by analyzing them? (P)

TV$_{10}$: Are the primary studies in your dataset published in a limited set of venues?

MA$_1$: Have you used snowballing? (P)

MA$_2$: Have you included grey literature (if this does not affect TV$_6$)? (P)

MA$_3$: Have you manually scanned selected venues to check if they publish articles related to your secondary study? (P)

TV$_{11}$: Do you expect to identify relationships in your dataset?

MA$_1$: Have you performed pilot data extraction to test the existence of relationships? (P)

TV$_{12}$: Does the quality of primary studies guarantee the validity of extracted data?

MA$_1$: Have you focused your search process on quality venues only? (P)

MA$_2$: Have you used article quality assessment as an inclusion criterion? (C)

MA$_3$: Have you assessed the validity of primary studies and their impact using statistics? (C)

TV$_{13}$: Is there data extraction bias in your study?

MA$_1$: Have you involved more than one researcher? (P)

MA$_2$: Have you identified experts' disagreement level with the kappa statistic? (P)

MA$_3$: Have you performed pilot data extraction to test agreement between researchers? (*Not applicable if* MA$_1$ *is no*) (P)

MA$_4$: Have you used experts *or* external reviewers' opinion in case of conflicts? (*Not applicable if* MA$_1$ *is no*) (C)

MA$_5$: Have you performed paper screening to cross-check data extraction? (P)

MA$_6$: Have you used a keywording of abstracts? (*Applicable only in mapping studies*) (P)

TV$_{14}$: Have you performed statistical analysis?

MA$_1$: Does your data extraction plan record quantitative data and if yes, does answering your research questions imply the use of statistics? (C)

TV$_{15}$: Have you selected a robust initial classification schema?

MA$_1$: Have you selected an existing initial classification schema? (P)

MA$_2$: Have you continuously updated the schema, until it becomes stable and classifies all primary studies in one or more classes? (C)

TV$_{16}$: Is your interpretation of the results subject to bias or is it as objective as possible?

MA$_1$: Have you performed pilot data analysis and interpretation? (P)

MA$_2$: Have you conducted reliability checks (e.g., post-SLR surveys with experts)? (C)

MA$_3$: Have you used a formal data synthesis method? (P)

MA$_4$: Have you performed sensitivity analysis? (P)

MA$_5$: Have you used the scientific quality of primary studies when drawing conclusions? (P)

Research Validity

TV_{17}: Is your process reliable/repeatable?

MA_1: Have more than one researcher been involved in the review process? (P)
MA_2: Have you made all gathered data publicly available? (C)
MA_3: Have you documented in detail the review process in a protocol? (P)
MA_4: Have you appropriately documented the details of conducting the review? (P)

TV_{18}: Have you chosen the correct research method?

MA_1: Have the authors discussed if the selected research method (SLR or SMS) fits the goals/research questions of the study, by advocating the purpose and scope of the methods? (C)
MA_2: Have you developed a protocol, monitored the process for deviations, and accurately reported any (if existed)? (P)

TV_{19}: Do the answers to your research questions guarantee the accomplishment of your study goal?

MA_1: Have the authors discussed *and* brainstormed on if the research questions holistically cover the goal of the study? (P)
MA_2: Is your study and research questions well-motivated? (P)
MA_3: Have you consulted target audience for setting your research goals? (P)

TV_{20}: Does your study have substantial related work, so that you can compare and discuss findings?

MA_1: Have the authors discussed *and* brainstormed to reach possible interpretations of the findings, due to the absence of related studies? (P)

TV_{21}: Were you familiar with the research field before performing the review?

MA_1: Have the authors exhaustively searched related work so as to: (a) familiarize with the field, (b) identify comparable studies, and (c) identify relevant publication venues and influential papers? (P)

TV_{22}: Are the results of your study generalizable?

MA_1: Do your findings comply with those of existing studies? (C)
MA_2: Have you used a broad search process without an initial starting date? (P)

As mentioned before, the main stakeholders of the checklist are the ***authors*** of a secondary study and the ***evaluator / reader*** of the study. For both stakeholder types, a possible use case scenario for the checklist is as follows:

[*STEP*1] The user is interested in evaluating the validity of a secondary study

[*STEP*2] The user asks the TV question, and if the answer suggest the existence of a threat (e.g., positive answer in TV_1 and negative in TV_{19}), then checks if there are any precautionary action (P) that can be taken. ***A threat to validity has been identified and needs to be reported.***

[*STEP*3] The user judges the effort required to perform the action (an estimate can be found in Section 6.2). If he/she decides to perform the action, the user checks if the threat is mitigated (an estimate of the fitness of each mitigation action is provided in Section 6.2). ***A mitigation action needs to be reported.*** If the threat is resolved, the user moves to the 2nd step and continues with the next TV question. ***The assessment of the outcome of the mitigation action needs to be reported.***

[*STEP*4] After the study is conducted, the corrective mitigation actions (C) for each TV question are visited. Step 3 is executed for each mitigation action.

[*STEP*5] The user goes through all the threats to validity questions and checks if at least one mitigation action has been performed.

6. Validation of classification schema and checklist

In this section, we present the validation of the proposed classification list and checklist, by applying the Delphi technique, with secondary study experts. This validation is necessary due to the nature of this study (i.e., the synthesized results provide guidelines for conducting future secondary studies); thus we want to make explicit the potential limitations and strengths of the classification schema and checklist, as identified by experts. In Section 6.1, we present the design of our empirical study, based on the guidelines provided by Runeson et al. [38], in Section 6.2 we report the results of the validation, and in Section 6.3 we discuss implications of this study to authors and readers of secondary studies.

6.1. Study design

Objectives and Research Questions: The goal of this study formulated in a GQM format [3] is: evaluate the proposed classification schema of threats to validity and the derived checklist, with respect to (a) the fitness of the threats to validity within their proposed categories, (b) the fitness of mitigation actions as a means of alleviating the corresponding threats, and (c) the effort required to apply each mitigation action, from the point of view of researchers in the context of empirical software engineering research. Based on this goal, we have formulated three research questions:

RQ_1: Are threats to validity correctly classified to the categories of the proposed schema?

RQ_2: Is the mapping between threats to validity and mitigation actions correct?

RQ_3: What is the effort required to apply each mitigation action?

RQ_1 aims at validating the first level of the classification schema depicted in Figs. 3a–c (classifying threats into categories). Similarly, RQ_2 aims at validating the relations at the second level of the schema (mapping threats to mitigation actions). Given the fact that some threats are mapped to several mitigation actions, RQ3 investigates the effort required to apply each mitigation action (RQ_3); this would be an interesting parameter when selecting among mitigation strategies for a particular threat. We note that the goal of this study is not to try to identify additional threats to validity but to validate the proposed classification schema and checklist.

Data Collection: To answer the aforementioned questions we decided to use experts' opinion by adopting a consensus method that is typically designed to combine the knowledge and experience of experts (e.g., [10,16,39,45]). We chose the Delphi technique [45] among the consensus methods because of the number of the participants we wanted to involve, and the time available to conduct the study. The Delphi technique is an iterative process that captures the opinions of different evaluators and at the same time records their levels of agreement [45]**.** As experts, we have selected a set of researchers with experience in secondary studies and empirical studies in general. The evaluators have been anonymized, but some demographics on their research experience are provided in Table 10 (based on the experts' pages in DBLP). The criteria that we have used in the participants selection process are the following: (a) all participants should be co-authors of at least 2 secondary studies, (b) all participants have published in the same high-quality venues, which we have in our sample, (c) all participants are senior academics—i.e., at least assistant professors; and (d) all participants work in different institutions.

The Delphi method [26] was applied with seven participants, which according to Verhaegen et al. [45] is an adequate number of experts. The number of iterations that we have performed is three (as also suggested in [45]). In the first round, the participants were given three questionnaires[8]:

- In the first questionnaire the participants were provided with the mapping between threat categories and specific threats. Each partic-

[8] http://se.uom.gr/wp-content/uploads/IST_material.zip

Table 10
Delphi Participants.

ID	Year of First Study (in general)	#Secondary Studies	#Primary Empirical Studies
P1	1986	4	31
P2	2006	4	13
P3	2002	4	9
P4	2001	8	1
P5	2010	11	0
P6	2012	2	2
P7	2009	2	1

ipant was asked to assign a Likert-scale value (1-Strongly Disagree, 2-Disagree, 3-Neutral, 4-Agree, and 5-Strongly Agree) that represents the fitness of the threat in the specific category. To guarantee the common understanding of categories and threats the participants were provided with the definitions described in this manuscript. Additionally, participants were asked to not take into consideration the perceived importance of the threat itself, but only its fitness to the category. As an alternative to this study setup one could argue in favor of questions that have to do with the extent to which the expert has used this threat, or if he/she would be willing to use it in future secondary studies. However, this setup was not considered, since such information could have easily been retrieved by exploring the threats to validity already reported in the approximately 40 secondary studies that the selected experts have published.

- The second questionnaire was similar to the first, with the difference that it mapped mitigation actions to specific threats to validity. The rest of the questionnaire setup was the same as in the first questionnaire.
- The third questionnaire listed all mitigation actions and asked the participants to assess the effort required to apply the mitigation action. The scale in this questionnaire ranged from 1-Very Low Effort to 5-Very High Effort.

During the 2nd and the 3rd iteration the participants were provided the mode score for each question from the previous round. Then, the participants were given the opportunity to revisit their answers, by considering the results of the previous round and changing their assessment for questions that they had second thoughts.

Data Analysis: Within each iteration (for each question), we calculated the mode score from all participants' responses and the frequencies of each value. As an acceptable level of agreement we have set frequencies higher than 40%. The results for each research question are visualized through bar charts and tables. Additionally, since the Likert scale allows only integer values, the mode value is also calculated so as to present the most popular answer among the participants.

6.2. Results of validation

The classification of threats to validity to specific categories (RQ1), and the results are summarized in Table 11. The first column of the figure denotes the categories introduced in Fig. 2, whereas the 2nd column the specific threats classified in these categories (see Section 5.1). The next six columns represent the frequencies of the answers in the Likert scale (i.e., the percentage that each range, from 1 to 5, received, compared the whole population), as obtained after the 3rd Delphi round. The green-shaded cells correspond to threats that the majority of experts at least Agree with their classification, whereas pink-shaded correspond to threats that most evaluators agree (or strongly agree) with their classification, but there was one/two objections (usually scored as neutral). The validity threats ratings range that receive the greater percentage are denoted with bold.

Regarding RQ2 and RQ3 (i.e., the fitness of mitigation actions to resolve specific threats to validity, and the effort required for their application), the results are graphically summarized in Figs. 5a–c. Each figure corresponds to one category of threats to validity (study selection, data validity, and research validity respectively). The height of the bar denotes the mode fitness (most popular score) of each mitigation action to alleviate the corresponding threat to validity, whereas the average effort is represented by the line. Optimally, a high bar with a low line denotes a suitable mitigation action that requires little effort. The threats to validity that each mitigation action resolves are denoted with the red boxes, labelled after the threat to validity. Based on Fig. 5a (regarding Study Selection validity), we can observe that the only mitigation action that is ranked as Neutral is "Systematic Voting" as a way to mitigate the issue of "Inefficient Selection of Inclusion/Exclusion Criteria". Nevertheless, for every threat of the Study Selection category there is at least one threat that is ranked with Strongly Agree. Among them, the least effort-intensive are "Comparison to a Gold Standard" for the Inadequacy of initial relevant publications identification and "Use of Summaries of Articles" for mitigating the "Inefficient handling of duplicate articles".

Finally, the findings of Fig. 5c, suggest that the mitigation actions for Research Validity threats are efficient in terms of fitness to resolve the problem, but on the other hand they are (even marginally) the most effort-intensive. This finding is intuitive since any corrective action at the process level is expected to be more time consuming, compared to activities focusing on the handling of particular primary studies.

6.3. Implications to authors & readers of secondary studies

We argue that the use of the provided classification schema (see Section 5.1) and checklist (see Section 5.2), in future secondary studies is expected to lead to the following benefits concerning both the readers and the authors of secondary studies:

- The authors of secondary studies can use the findings reported in this work to comprehensively identify potential threats to the validity of their studies and reuse mitigation actions. This will allow the transfer of knowledge among researchers and hinder the "reinvention" of threats to validity and mitigation actions for every secondary study.
- The reporting of threats to validity will be enhanced. In particular, the reporting of secondary studies threats to validity can be structured based on the categorization that is provided in our classification schema.
- The readers of the secondary studies will be able to uniformly interpret the results of the studies, and will be able to compare the quality and credibility of secondary studies.
- The readers (or potential reviewers of the studies, prior to their publication) can use the proposed classification schema and checklist to assess the validity of the study.

As an interesting future work direction we note that since the effort of applying the mitigation action and its benefit (i.e., fitness for resolving the threat) have been assessed in this study, it they can be used in a version of a cost-benefit analysis for trade-off management.

Table 11
Validity of Classifying Threats to Categories.

		1	2	3	4	5	N/A
Study Selection	Inadequacy of initial relevant publications identification	0%	0%	0%	**42.9%**	**42.9%**	14.3%
	Limited number of journals and conferences	0%	0%	0%	28,6%	**57,1%**	14.3%
	Missing non-English papers	0%	0%	28.6%	**42.9%**	14.3%	14.3%
	Paper not accessible in a digital library	0%	0%	0%	28,6%	**57,1%**	14.3%
	Inefficient handling of duplicate articles	0%	0%	28,6%	28,6%	28,6%	14.3%
	Inclusion / exclusion of grey literature	0%	0%	0%	**42.9%**	**42.9%**	14.3%
	Insufficient study inclusion / exclusion criteria	0%	0%	0%	28.6%	**71.4%**	0%
Data Validity	Small sample size or heterogeneous primary studies	0%	0%	0%	28.6%	**57,1%**	14.3%
	The chosen variables to be extracted cannot answer the RQs	0%	0%	0%	0%	**100%**	0%
	Primary studies are published in a limited number of venues	0%	0%	28.6%	**42.9%**	14.3%	14.3%
	The obtained dataset lacks relationships	0%	0%	14.3%	**71.4%**	0%	14.3%
	Low validity of primary studies	0%	0%	0%	42.9%	**57.1%**	0%
	Data extraction is biased	0%	0%	0%	28.6%	**71.4%**	0%
	No statistical analysis of the dataset	0%	0%	0%	**42.9%**	**42.9%**	14.3%
	The selection of classification schema is biased	0%	0%	0%	42.9%	**57.1%**	0%
	The interpretation of results is not objective	0%	0%	0%	14.3%	**85.7%**	0%
Research Validity	Lack of repeatability	0%	0%	0%	14.3%	**85.7%**	0%
	A not fitting research method has been selected	0%	0%	0%	14.3%	**85.7%**	0%
	Answering the RQs cannot fulfill the goal	0%	0%	0%	14.3%	**85.7%**	0%
	Lack of comparable studies	0%	0%	14.3%	14.3%	**57.1%**	14.3%
	Researchers are not familiar with the research field	0%	0%	0%	28.6%	**57.1%**	14.3%
	Lack of generalizability	0%	0%	0%	28.6%	**71.4%**	0%

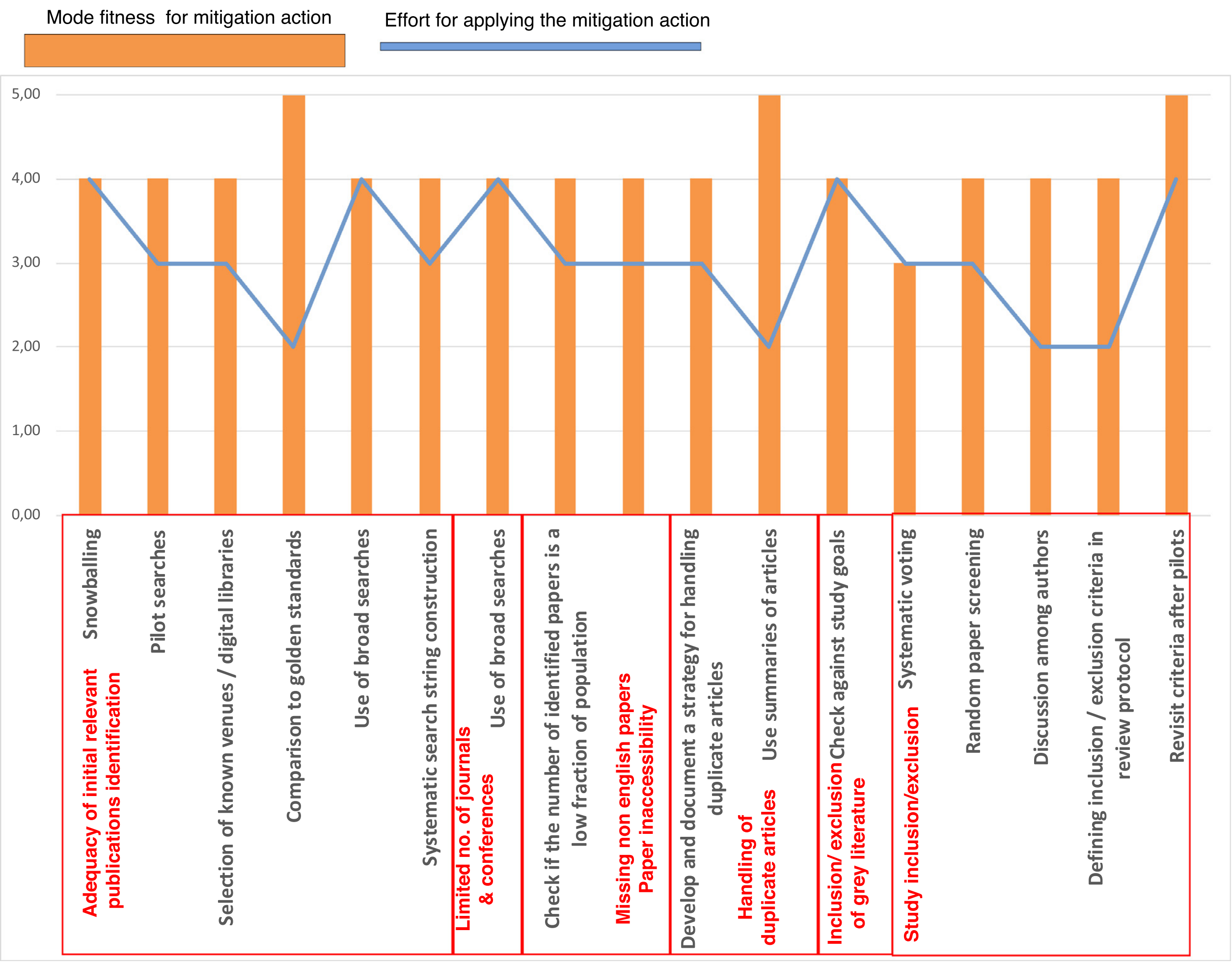

Fig. 5a. Mitigation Actions for Study Selection Threats to Validity.

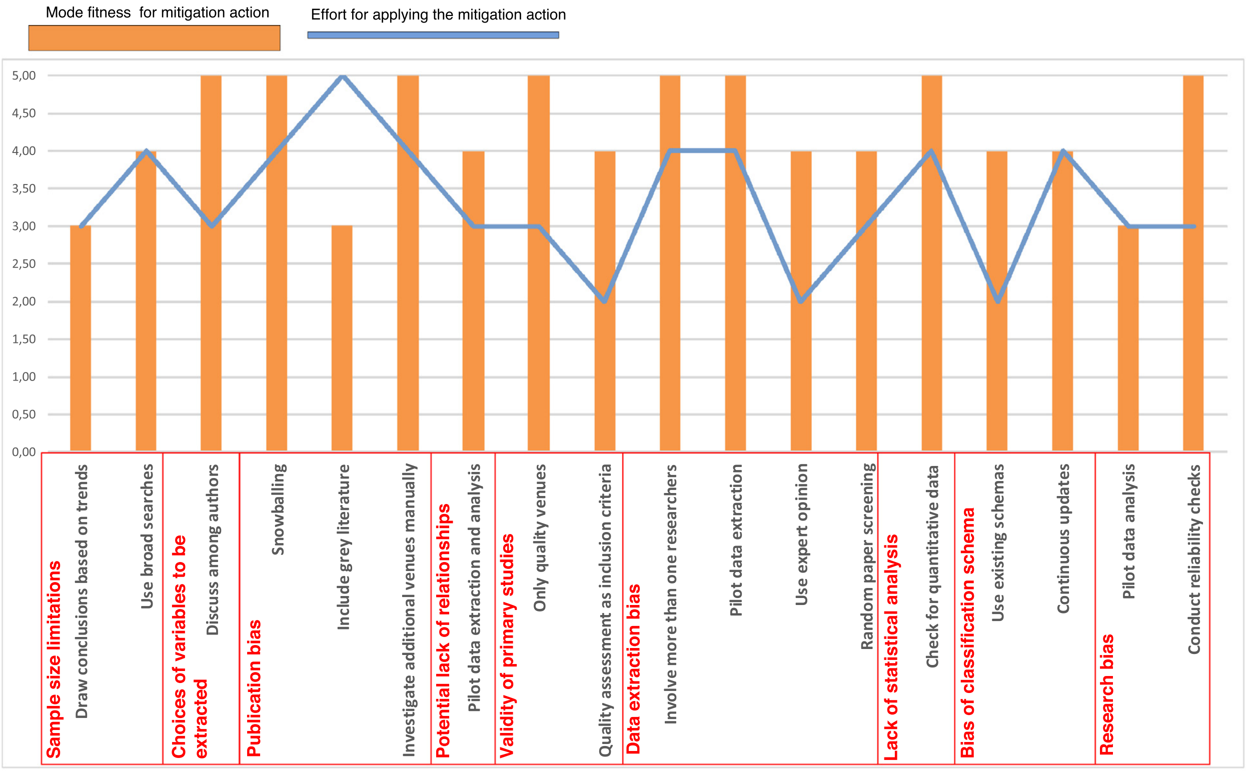

Fig. 5b. Mitigation Actions for Data Validity Threats.

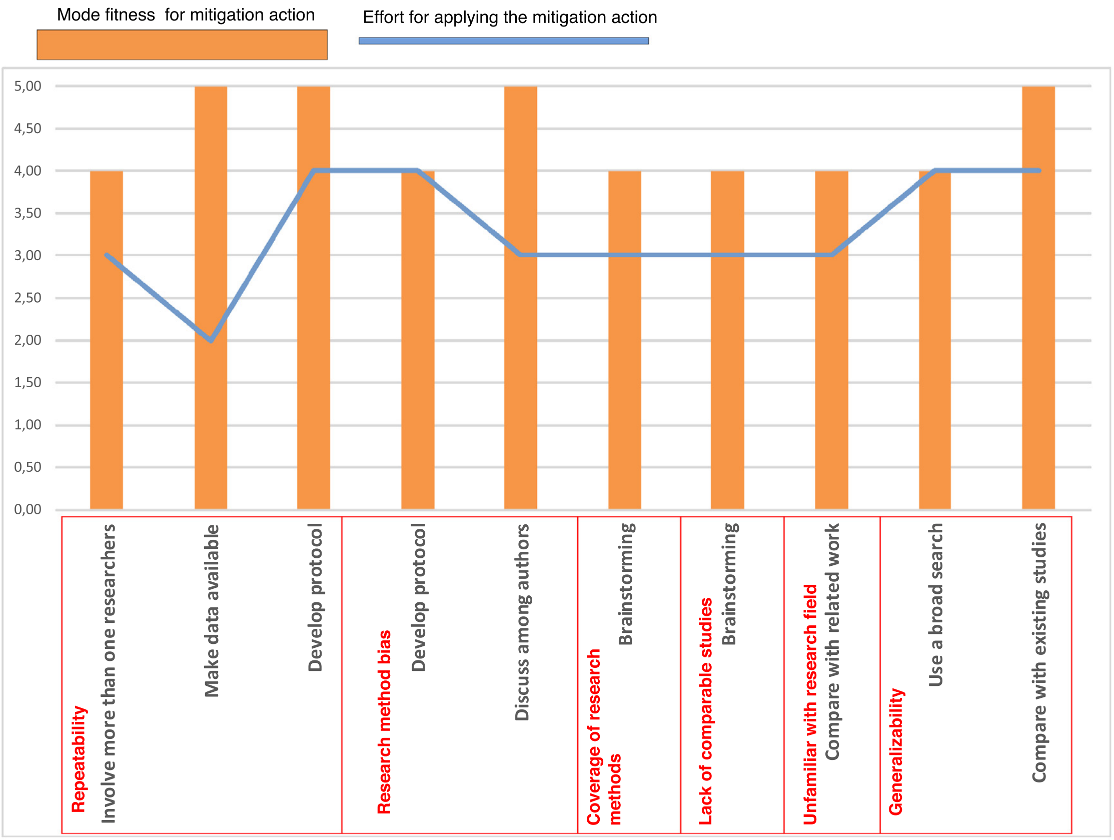

Fig. 5c. Mitigation Actions for Research Validity Threats.

7. Threats to validity

In this section we present the threats to validity that we have identified for this tertiary study. In order for this section to act as a proof of concept for the classification proposed in this work, we structure this section, based on the checklist provided in Section 5.2. Specifically, in Section 7.1, we report threats to validity related to study selection (TV_1-TV_7), in Section 7.2, we report threats related to data validity (TV_8-TV_{14}), and in Section 7.3, we report threats related to research validity (TV_{15}-TV_{22}).

7.1. Study selection validity

Study selection validity is recognized as the major threat in secondary studies during the early phases of the research. In this case, in order to ensure that our searching process has *adequately identified all relevant studies* (TV_1), the secondary studies that have been selected for inclusion have been carefully chosen following a well-defined protocol based on strict guidelines [18]. The identification procedure consisted of an automated search through the search engines of the most-known DLs for articles published in well-established journals and conferences. The search strings ("survey", "literature review", "mapping study", "mapping studies", "systematic review", "systematic mapping") that were used are quite broad, since we only included the name of the investigated research method, aiming to retrieve the maximum number of relevant studies. However, studies that adopted different terminology than the most established one might have been excluded. Nevertheless, we note that our study focused only on research efforts that are aware of the processes for conducting secondary studies and use established guidelines for this reason. To mitigate the risk of losing relevant studies we validated our set of secondary studies by cross-checking them against papers in other tertiary studies (serving as a gold standard). The results of this process suggested that we have been able to obtain approximately 95% of secondary studies that are referenced in other tertiary studies.

After the set of secondary studies has been obtained, we proceeded to the article inclusion/exclusion phase, which is threatened by the possibility to *exclude some relevant articles (TV_7)*. To mitigate this threat, two researchers have been involved in this process, discussing any possible conflicts. On the completion of this process, a third researcher was randomly screening the selection of articles for inclusion. Also, the inclusion/exclusion criteria have been extensively discussed among the authors, so as to guarantee their clarity and prohibit misinterpretations. Furthermore, from our searching space we have *excluded grey literature* (TV_6), since the goal of the study was imposing the use of only a limited number of journals and conferences that would guarantee the quality of the obtained papers.

Additionally, although we have not identified any *duplicate articles* (TV_5), our research protocol dictated that we check for duplicated articles, based on the abstract. Upon identification, the most extensive study would be retained. Also, our study is not suffering from the *missing non-English papers* (TV_3) and the *papers published in a limited number of journals and conferences* (TV_2), since our search process was aiming to a large number of publication venues all publishing papers only in English. Finally, we have been able to *access all publications* (TV_4) that we were interested in, since our research institutes provide us access to the used DLs.

7.2. Data validity

Regarding data validity, the main threat is related to *data extraction bias (TV_{13})*. In this phase, all relevant data were extracted and recorded manually by the second author. Obviously, and since this procedure inserted some subjectivity we mitigated this threat since two researchers that worked in-pair further inspected and refined the collected data, revalidating them. After this procedure the results were discussed among all researchers and any conflicts have been resolved.

Additionally, *publication bias* (TV_{10}) is present in our results since most of the secondary studies explored come from two dominating journals in the area of SE (IST, JSS). Since the quality of the results would be jeopardized if "grey literature" or non-indexed publication titles were included there was no option but to include only the venues presented in Table 3. Nevertheless, we believe that the obtained data points are not influenced by a small group of people, but from the software engineering community as a whole, since they stand among the first selections of publication venues for high-quality research in the world-wide community.

One of the aims of the proposed classification schema (apart from the provision of a common vocabulary to authors and readers) is to alleviate the aforementioned problem. Although we believe that the current classification schema improves the orthogonality among threat categories: (a) there are still some grey-zone threats, and (b) there is a cause-effect relationship between threats (see Section 5.1). Although we have not used an *initial classification schema* (TV_{15}) for our review (since existing ones were not matching—see Section 5.1), we have continuously iterated on developing a new one. Nevertheless, we need to note that as a starting point, we have used the phases of the systematic literature review process, see Kitchenham et al. [11].

Finally, our tertiary study is not threatened by the following threats: (a) *small sample size (TV_8)*—we have been able to retrieve approx. 100 articles, (b) *lack of relationships (TV_{11})*—our study was not aiming to identify any relationships among data, but only to classify and synthesize, (c) *low quality of primary studies (TV_{12})*—since the involved studies have been published only in top software engineering venues, and (d) *selection of variables to be extracted (TV_9)*—the straightforward research questions of our study have not raised any conflicts in the discussions among authors on which variables should be extracted. Finally, the study does not *lack the use of statistical analysis* (TV_{14}), since x^2 testing and linear regression have been performed to answer RQ_1. The nature of RQ_2 and RQ_3 led us to the decision to only perform some basic statistical analysis (descriptive), since no hypothesis testing was necessary. Finally, to mitigate the *researchers' bias in data interpretation and analysis* the authors have discussed the threats to validity's classification and clustering *(TV_{16})*.

7.3. Research validity

Concerning research validity, we have been able to exclude two possible threats to validity due, to the nature of our study. First, the authors are highly *familiar with secondary studies* (TV_{21}), since they have been involved in a large number as authors and reviewers. Therefore, no mitigation actions were necessary. Furthermore, we believe that the followed review process ensures the *reliability* (TV_{17}) and safe replication of our study. First, all important decisions in our review planning have been thoroughly documented in this manuscript (see Section 3) are can be easily reproduced by other researchers. Second, the fact that the data extraction was based on the opinion of three researchers can to some extent guarantee the elimination of bias, making the dataset reliable. Third, all extracted data have been made publicly available, so as to enable comparison of results[3]. Nevertheless, some threats to research validity have been identified and mitigated. First, through discussion among the authors we have set four *research questions that accurately and holistically map to the set goal (TV_{19})*. This is clearly depicted by the mapping of each research question to the research sub-goals/objectives (see Section 3.1). Second, in the literature we have been able to identify a substantial amount of *related works that can be used for comparison* (TV_{20}) to our results. In particular, for this reason we used related studies from software engineering and medical literature. Third, the *selection of the research method* (TV_{18}) is adequate for the goal of this study (since plenty of synthesis was required) and no deviations from the guidelines have been performed.

Concerning *generalizability* (TV_{22}), we can claim that our results comply both with existing literature and with common sense (i.e., secondary studies are less mature than surveys, case studies, and experiments). To ensure the generalizability of our results we have examined a wide range of studies, from all subfields of software engineering, without any focus on some specific activity (e.g., maintenance, architecture, etc.). Therefore, we believe that our results are generalizable to good quality papers in the software engineering domain, but not necessarily to grey literature, other venues, and other disciplines.

8. Conclusion

In the last decade, secondary studies (i.e., systematic literature reviews and mapping studies) have emerged as a popular research methodology for summarizing existing literature. Despite their popularity and the thorough guidelines for conducting them, the research state-of-the-art lacks support on how to identify, report and mitigate threats to validity for secondary studies. To alleviate this problem we have conducted a tertiary study on software engineering research corpus, i.e., a literature review of literature reviews. The final goal of this tertiary study was to develop a classification schema with three levels: (a) threats categories, (b) specific threats, and (c) mitigation actions.

The results of the study suggested that there are three main categories of threats: (a) threats to study selection, (b) threats to data collection, and (c) threats to research validity. Each category includes approximately ten specific threats to validity, which can be mitigated with at least one action. To facilitate the easy application of this classification schema, a checklist with questions that can guide the authors and readers of secondary studies in assessing study validity has been provided. In particular, on the one hand, authors of secondary studies can use the checklist for identifying threats to validity and get access to a list of possible mitigation actions. On the other hand, the readers of secondary studies can use the checklist to evaluate the validity of the obtained results. To validate the obtained results, we empirical assessed them by employing the Delphi method with experts evaluating the fitness of mitigation actions as a means of alleviating the corresponding threats and provided an estimate of the effort required to apply each mitigation action.

Appendix A. Threats to Validity Description

Threats to Study Selection Validity

Name	Description
Selection of arbitrary starting year	The selection of a specific year as a starting point for performing the search process can lead to missing studies prior to that date.
Search engine inefficiencies	Problems of the DLs search engines (e.g., SpringerLink cannot perform a search based only on the abstract of manuscripts). This can lead to missing studies, or deriving a large corpus of papers for filtering.
Limited number of journals & conferences	A limited number of publication venues in which primary studies can be published suggest a narrow scope of the secondary study. This will probably lead to obtaining a low number of primary studies.
Missing non-English papers	Exploring studies written in a specific language can lead to the omission of important studies (or number of studies) written in other languages.
Papers inaccessibility	Papers whose full-text is not available cannot be processed. If this number is large, the set of studies might be limited / not representative.
Handling of duplicate articles	Some early versions of a study may be published in a conference, and an extended one in a journal. Duplicate studies should be identified and handled, so that the study set, does not contain duplicate information.
Inclusion/Exclusion of Grey literature	Based on the goal of the study, including or excluding grey literature can pose a threat. For example, grey literature should be considered in Multi-Vocal Literature Reviews (MLRs), in which practitioners' view should be examined.

Threats to Data Validity

Name	Description
Small sample size	A small sample threatens the validity of the dataset, since results may be: (a) prone to bias (data might come from a small community), (b) not statistically significant, and (c) not safe to generalize.
Heterogeneity of primary studies	Data that are highly heterogeneous are not easy / safe to synthesize, since such a process is prone to involve a high degree of subjectivity.
Choices of variables to be extracted	The variables that have been chosen to be extracted might threaten the validity of the results, since they might not fit for answering the research questions. Additionally, they are prone to researchers' bias.
Potential lack of relationships	Examining data that lack relations might hinder reaching a conclusion.
Data extraction inaccuracies	Data analysis might not be carefully performed, or might not follow strict guidelines. For example, the same concept might be inconsistently classified into two primary studies. This leads to inaccuracies in the dataset.
Unverified data extraction	Data extraction items that are not verified by external reviewers, or have not been subject to internal review.
Miss-classification of primary studies	This threat is valid for secondary studies that aim at developing a classification schema (usually mapping studies). This threat can occur if primary studies are incorrectly or inconsistently classified in a specific class.
Lack of statistical analysis	In some designs it is not possible to perform statistical analysis. For example, in cases that all extracted data items are categorical.
Construction of attribute framework	When we define a set of possible values for the attributes (i.e., variables) that are used to characterize each primary study, we construct an attribute framework. If the selected values are not discriminative and comprehensive then the data extraction can result to an insufficient dataset

Threats to Research Process Validity

Name	Description
Chosen research method	Mapping studies and literature reviews are designed to serve different goals and scopes. The selection of a specific research method might not fit the goals, the scope, or the context of the performed secondary study.
Review process deviations	In some cases researchers choose to deviate from the guidelines offered by the research method. Such deviations (e.g., not performing the keywording of abstracts step in a mapping study, although the guidelines of Petersen [22] are used) threaten the validity of the study, since some important aspects might be compromised.
Lack of comparable studies	Some secondary studies lack comparable related work (i.e., other secondary studies or primary studies). In this case there is no possibility of comparing the results to existing literature, or intuitively validate them.
Unfamiliarity to the research field	In some cases secondary studies are performed by non-expert researchers. The lack of knowledge in the domain can lead to undesired consequences, such as: omission of well-known studies in the field, limited synthesis capacity, etc.

Appendix B. List of Studies

[S1] Abdelmaboud, A., Jawawi, D. N., Ghani, I., Elsafi, A., and Kitchenham, B. Quality of service approaches in cloud computing: A systematic mapping study. Journal of Systems and Software, 101, 159–179, 2015.

[S2] Abelein, U., and Paech, B. Understanding the Influence of User Participation and Involvement on System Success: a Systematic Mapping Study. Empirical Software Engineering, 20(1), 28–81, 2015.

[S3] Achimugu, P., Selamat, A., Ibrahim, R., and Mahrin, M. N. A systematic literature review of software requirements prioritization research. Information and Software Technology, 56(6), 568–585, 2014.

[S4] Afzal, W., Torkar, R., and Feldt, R. A systematic review of search-based testing for non-functional system properties. Information and Software Technology, 51(6), 957–976, 2009.

[S5] Ahmad, A., and Babar, M. A. Software architectures for robotic systems: A systematic mapping study. Journal of Systems and Software, 122, 16–39, 2016.

[S6] Al Dallal, J. Identifying refactoring opportunities in object-oriented code: A systematic literature review. Information and Software Technology, 58, 231–249, 2015.

[S7] Al-Baik, O., and Miller, J. The kanban approach, between agility and leanness: a systematic review. Empirical Software Engineering, 20(6), 1861–1897, 2015.

[S8] Aleti, A., Buhnova, B., Grunske, L., Koziolek, A., and Meedeniya, I. (2013, #may#). Software Architecture Optimization Methods: A Systematic Literature Review. IEEE Transactions on Software Engineering, 39(5), 658–683, 2013.

[S9] Ali, M. S., Ali Babar, M., Chen, L., and Stol, K.-J. A systematic review of comparative evidence of aspect-oriented programming. Information and Software Technology, 52(9), 871–887, 2010.

[S10] Ali, S., Briand, L., Hemmati, H., and Panesar-Walawege, R. A Systematic Review of the Application and Empirical Investigation of Search-Based Test Case Generation. IEEE Transactions on Software Engineering, 36(6), 742–762, 2010.

[S11] Alves, N. S., Mendes, T. S., de Mendonsa, M. G., Spinola, R. O., Shull, F., and Seaman, C. Identification and management of technical debt: A systematic mapping study. Information and Software Technology, 70, 100–121, 2016.

[S12] Alves, V., Niu, N., Alves, C., and Valença, G. Requirements engineering for software product lines: A systematic literature review. Information and Software Technology, 52(8), 806–820, 2010.

[S13] Ameller, D., Burgués, X., Collell, O., Costal, D., Franch, X., and Papazoglou, M. P. Development of service-oriented architectures using model-driven development: A mapping study. Information and Software Technology, 62, 42–66, 2015.

[S14] Ampatzoglou, A., Ampatzoglou, A., Chatzigeorgiou, A., and Avgeriou, P. The Financial Aspect of Managing Technical Debt: A Systematic Literature Review. Information and Software Technology, 2015.

[S15] Ampatzoglou, A., and Stamelos, I. Software engineering research for computer games: A systematic review. Information and Software Technology, 52(9), 888–901, 2010.

[S16] Ampatzoglou, A., Charalampidou, S., and Stamelos, I. Research state of the art on GoF design patterns: A mapping study. Journal of Systems and Software, 86(7), 1945–1964, 2013.

[S17] Anh, N.-D., Cruzes, D. S., and Conradi, R. Dispersion, Coordination and Performance in Global Software Teams: A Systematic Review. Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 129–138, New York, NY, USA: ACM, 2012.

[S18] Anjum, M., and Budgen, D., "A mapping study of the definitions used for Service Oriented Architecture." 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012).

[S19] Arias, T. B., Spek, P. v., and Avgeriou, P. A practice-driven systematic review of dependency analysis solutions. Empirical Software Engineering, 16(5), 544–586, 2011.

[S20] Badampudi, D., Wohlin, C., and Petersen, K. Software component decision-making: In-house, OSS, COTS or outsourcing - A systematic literature review. Journal of Systems and Software, 121, 105–124, 2016.

[S21] Bakar, N. H., Kasirun, Z. M., and Salleh, N. Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review. Journal of Systems and Software, 106, 132–149, 2015.

[S22] Bano, M., & Zowghi, D. (2013, April). User involvement in software development and system success: a systematic literature review. In Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering (pp. 125–130). ACM.

[S23] Bano, M., Imtiaz, S., Ikram, N., Niazi, M., & Usman, M. Causes of requirement change-A systematic literature review. In 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012).

[S24] Barreiros, E., Almeida, A., Saraiva, J., and Soares, S. A Systematic Mapping Study on Software Engineering Testbeds. 2011 International Symposium on Empirical Software Engineering and Measurement (ESEM), 107–116, 2011.

[S25] Bissi, W., Serra Seca Neto, A. G., and Emer, M. C. The effects of test driven development on internal quality, external quality and productivity: A systematic review. Information and Software Technology, 74, 45–54, 2016.

[S26] Bjornson, F. O., and Dingsoyr, T. Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used. Information and Software Technology, 50(11), 1055–1068, 2018.

[S27] Borg, M., Runeson, P., and Ardö, A. Recovering from a decade: a systematic mapping of information retrieval approaches to software traceability. Empirical Software Engineering, 19(6), 1565–1616, 2013.

[S28] Borges, A., Ferreira, W., Barreiros, E., Almeida, A., Fonseca, L., Teixeira, E & Soares, S. (2015, April). Support mechanisms to conduct empirical studies in software engineering: a systematic mapping study. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (p. 22). ACM.

[S29] Breivold, H. P., Crnkovic, I., and Larsson, M. A systematic review of software architecture evolution research. Information and Software Technology, 54(1), 16–40, 2012.

[S30] Budgen, D., Burn, A. J., Brereton, O. P., Kitchenham, B. A., and Pretorius, R. Empirical evidence about the UML: a systematic literature review. Software: Practice and Experience, 41(4), 363–392, 2011.

[S31] Campanelli, A. S., and Parreiras, F. S. Agile methods tailoring: A systematic literature review. Journal of Systems and Software, 110, 85–100, 2015.

[S32] Carroll, C., Falessi, D., Forney, V., Frances, A., Izurieta, C., and Seaman, C. A Mapping Study of Software Causal Factors for Improving Maintenance. 2015 ACMIEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 1–4, 2015.

[S33] Chang, B.-M., and Choi, K. A review on exception analysis. Information and Software Technology, 77, 1–16, 2016.

[S34] Chauhan, M. A., Babar, M. A., and Benatallah, B. Architecting cloud-enabled systems: a systematic survey of challenges and solutions: A Systematic Survey of Cloud Architecting Challenges and Solutions. Software: Practice and Experience, 2016.

[S35] Cornelissen, B., Zaidman, A., van Deursen, A., Moonen, L., and Koschke, R. A Systematic Survey of Program Comprehension through Dynamic Analysis. IEEE Transactions on Software Engineering, 35(5), 684–702, 2009.

[S36] da Mota Silveira Neto, P. A., Carmo Machado, I. d., McGregor, J. D., de Almeida, E. S., and de Lemos Meira, S. R. A systematic mapping study of software product lines testing. Information and Software Technology, 53(5), 407–423, 2011.

[S37] da Silva, I. F., da Mota Silveira Neto, P. A., O'Leary, P., de Almeida, E. S., and de Lemos Meira, S. R. Agile software product lines: a systematic mapping study. Software: Practice and Experience, 41(8), 899–920, 2011.

[S38] Dasanayake, S., Markkula, J., & Oivo, M. (2014, May). Concerns in software development: a systematic mapping study. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (p. 21). ACM.

[S39] de Magalhães, C. V., da Silva, F. Q., & Santos, R. E. (2014, May). Investigations about replication of empirical studies in software engineering: preliminary findings from a mapping study. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (p. 37). ACM.

[S40] de MagalhΓ£es, C. V., da Silva, F. Q., Santos, R. E., and Suassuna, M. Investigations about replication of empirical studies in software engineering: A systematic mapping study. Information and Software Technology, 64, 76–101, 2015.

[S41] Diebold, P., & Dahlem, M. (2014, May). Agile practices in practice: a mapping study. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (p. 30). ACM.

[S42] Dieste, O., and Juristo, N. Systematic review and aggregation of empirical studies on elicitation techniques. IEEE Transactions on Software Engineering, 37(2), 283–304, 2011.

[S43] Dikert, K., Paasivaara, M., and Lassenius, C. Challenges and success factors for large-scale agile transformations: A systematic literature review. Journal of Systems and Software, 119, 87–108, 2016.

[S44] Ding, W., Liang, P., Tang, A., and van Vliet, H. Knowledge-based approaches in software documentation: A systematic literature review. Information and Software Technology, 56(6), 545–567, 2014.

[S45] Dominguez, E., Pérez, B., Rubio, Á. L., and Zapata, M. A. A systematic review of code generation proposals from state machine specifications. Information and Software Technology, 54(10), 1045–1066, 2012.

[S46] Dov gan, S., Betin-Can, A., and Garousi, V. Web application testing: A systematic literature review. Journal of Systems and Software, 91, 174–201, 2014.

[S47] Elberzhager, F., Münch, J., and Nha, V. T. A systematic mapping study on the combination of static and dynamic quality assurance techniques. Information and Software Technology, 54(1), 1–15, 2012.

[S48] Elberzhager, F., Rosbach, A., Münch, J., and Eschbach, R. Reducing test effort: A systematic mapping study on existing approaches. Information and Software Technology, 54(10), 1092–1106, 2012.

[S49] Engström, E., and Runeson, P. Software product line testing – A systematic mapping study. Information and Software Technology, 53(1), 2–13, 2011.

[S50] Engström, E., Runeson, P., and Skoglund, M. A systematic review on regression test selection techniques. Information and Software Technology, 52(1), 14–30, 2010.

[S51] Engström, E., Skoglund, M., and Runeson, P. Empirical Evaluations of Regression Test Selection Techniques: A Systematic Review. Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 22–31. New York, NY, USA: ACM, 2008.

[S52] Febrero, F., Calero, C., and Moraga, M. Á. A Systematic Mapping Study of Software Reliability Modeling. Information and Software Technology, 56(8), 839–849, 2014.

[S53] Fernandes, E., Oliveira, J., Vale, G., Paiva, T., & Figueiredo, E. (2016, June). A review-based comparative study of bad smell detection tools. In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering (p. 18). ACM.

[S54] Fernandez, A., Insfran, E., and Abrahão, S. Usability evaluation methods for the web: A systematic mapping study. Information and Software Technology, 53(8), 789–817, 2011.

[S55] Fernández-Sáez, A. M., Genero, M., and Chaudron, M. R. Empirical studies concerning the maintenance of UML diagrams and their use in the maintenance of code: A systematic mapping study. Information and Software Technology, 55(7), 1119–1142, 2013.

[S56] Galster, M., Weyns, D., Tofan, D., Michalik, B., and Avgeriou, P. Variability in Software Systems—A Systematic Literature Review. IEEE Transactions on Software Engineering, 40(3), 282–306, 2014.

[S57] Garcia, S., Romero, O., and RaventΓ³s, R. (2016). DSS from an RE Perspective: A systematic mapping. Journal of Systems and Software, 117, 488–507, 2016.

[S58] Garousi, V., Amannejad, Y., and Betin Can, A. Software test-code engineering: A systematic mapping. Information and Software Technology, 58, 123–147.

[S59] Garousi, V., and Mantyla M. V. When and what to automate in software testing? A multi-vocal literature review. Information and Software Technology, 76, 92–117, 2013.

[S60] Garousi, V., and Mantyla, M. V. A systematic literature review of literature reviews in software testing. Information and Software Technology, 80, 195–216.

[S61] Garousi, V., Mesbah, A., Betin-Can, A., and Mirshokraie, S. A systematic mapping study of web application testing. Information and Software Technology, 55(8), 1374–1396, 2013.

[S62] Garousi, V., Petersen, K., and Ozkan, B. Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review. Information and Software Technology, 79, 106–127, 2016.

[S63] Gasparic, M., and Janes, A. What recommendation systems for software engineering recommend: A systematic literature review. Journal of Systems and Software, 113, 101–113, 2016.

[S64] Gholami, M. F., Daneshgar, F., Low, G., and Beydoun, G. Cloud migration process*β*€" A survey, evaluation framework, and open challenges. Journal of Systems and Software, 120, 31–69, 2016.

[S65] González, C. A., and Cabot, J. Formal verification of static software models in MDE: A systematic review. Information and Software Technology, 56(8), 821–838, 2014.

[S66] Gonzalez-Landron-de-Guevara, F., Fernandez-Diego, M., and Lokan, C. The usage of ISBSG data fields in software effort estimation: A systematic mapping study. Journal of Systems and Software, 113, 188–215, 2016.

[S67] Guinea, A. S., Nain, G., and Traon, Y. L. A systematic review on the engineering of software for ubiquitous systems. Journal of Systems and Software, 118, 251–276, 2016.

[S68] Hall, T., Baddoo, N., Beecham, S., Robinson, H., and Sharp, H. A Systematic Review of Theory Use in Studies Investigating the Motivations of Software Engineers. ACM Trans. Softw. Eng. Methodology, 18(3), 10:1–10:29, 2009.

[S69] Hall, T., Beecham, S., Bowes, D., Gray, D., and Counsell, S. A Systematic Literature Review on Fault Prediction Performance in Software Engineering. IEEE Transactions on Software Engineering, 38(6), 1276–1304, 2012.

[S70] Haselberger, D. A literature-based framework of performance-related leadership interactions in ICT project teams. Information and Software Technology, 70, 1–17, 2016.

[S71] Häser, F., Felderer, M., & Breu, R. (2014, May). Software paradigms, assessment types and non-functional requirements in model-based integration testing: a systematic literature review. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (p. 29). ACM.

[S72] Heckman, S., and Williams, L. A systematic literature review of actionable alert identification techniques for automated static code analysis. Information and Software Technology, 53(4), 363–387, 2011.

[S73] Holl, G., Grünbacher, P., and Rabiser, R. A systematic review and an expert survey on capabilities supporting multi product lines. Information and Software Technology, 54(8), 828–852, 2012.

[S74] Idri, A., Amazal, F. a., and Abran, A. Analogy-based software development effort estimation: A systematic mapping and review. Information and Software Technology, 58, 206–230, 2015.

[S75] Idri, A., Hosni, M., and Abran, A. Systematic literature review of ensemble effort estimation. Journal of Systems and Software, 118, 151–175, 2016.

[S76] Irshad, M., Torkar, R., Petersen, K., & Afzal, W. (2016, June). Capturing cost avoidance through reuse: systematic literature review and industrial evaluation. In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering (p. 35). ACM.

[S77] Jabangwe, R., Borstler, J., Smite, D., and Wohlin, C. Empirical evidence on the link between object-oriented measures and external quality attributes: a systematic literature review. Empirical Software Engineering, 20(3), 640–693, 2015.

[S78] Jia, C., Cai, Y., Yu, Y. T., and Tse, T. H. 5W + 1H pattern: A perspective of systematic mapping studies and a case study on cloud software testing. Journal of Systems and Software, 116, 206–219, 2016.

[S79] Jorgensen, M., and Shepperd, M. A Systematic Review of Software Development Cost Estimation Studies. IEEE Transactions on Software Engineering, 33(1), 33–53, 2007.

[S80] Kabbedijk, J., Bezemer, C.-P., Jansen, S., and Zaidman, A. Defining multi-tenancy: A systematic mapping study on the academic and the industrial perspective. Journal of Systems and Software, 100, 139–148, 2015.

[S81] Khan, S. U., Niazi, M., and Ahmad, R. Barriers in the selection of offshore software development outsourcing vendors: An exploratory study using a systematic literature review. Information and Software Technology, 53(7), 693–706, 2011.

[S82] Khan, S. U., Niazi, M., and Ahmad, R. Factors influencing clients in the selection of offshore software outsourcing vendors: An exploratory study using a systematic literature review. Journal of Systems and Software, 84(4), 686–699, 2011.

[S83] Khurum, M., and Gorschek, T. A systematic review of domain analysis solutions for product lines. Journal of Systems and Software, 82(12), 1982–2003. 2009.

[S84] Kitchenham, B., Mendes, E., and Travassos, G. H. Cross versus Within-Company Cost Estimation Studies: A Systematic Review. IEEE Transactions on Software Engineering, 33(5), 316–329, 2007.

[S85] Kosar, T., Bohra, S., and Mernik, M. Domain-Specific Languages: A Systematic Mapping Study. Information and Software Technology, 71, 77–91, 2016.

[S86] Kuhrmann, M., Fernández, D. M., & Tiessler, M. (2013, April). "A mapping study on method engineering: first results" . In Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering (pp. 165–170). ACM.

[S87] Kupiainen, E., Mäntylä, M. V., and Itkonen, J. Using metrics in Agile and Lean Software Development – A systematic literature review of industrial studies. Information and Software Technology, 62, 143–163, 2015.

[S88] Kusumo, D. S., Staples, M., Zhu, L., Zhang, H., & Jeffery, R. (2012, May). Risks of off-the-shelf-based software acquisition and development: A systematic mapping study and a survey. In 16th International Conference on the Evaluation & Assessment in Software Engineering (EASE 2012), (pp. 233–242). IET.

[S89] Lenberg, P., Feldt, R., and Wallgren, L. G. Behavioral software engineering: A definition and systematic literature review. Journal of Systems and Software, 107, 15–37, 2015.

[S90] Li, J., Zhang, H., Zhu, L., Jeffery, R., Wang, Q., & Li, M. (2012, May). Preliminary results of a systematic review on requirements evolution. In 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012), (pp. 12–21). IET.

[S91] Li, Z., Avgeriou, P., and Liang, P. A systematic mapping study on technical debt and its management. Journal of Systems and Software, 101, 193–220, 2015.

[S92] Li, Z., Liang, P., and Avgeriou, P. Application of knowledge-based approaches in software architecture: A systematic mapping study. Information and Software Technology, 55(5), 777–794, 2013.

[S93] Li, Z., Zhang, H., O' Brien, L., Cai, R., and Flint, S. On evaluating commercial Cloud services: A systematic review. Journal of Systems and Software, 86(9), 2371–2393, 2013.

[S94] Liu, G., Rong, G., Zhang, H., & Shan, Q. (2015, April). The adoption of capture-recapture in software engineering: a systematic literature review. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (p. 15). ACM.

[S95] Lopez-Herrejon, R. E., Linsbauer, L., and Egyed, A. A systematic mapping study of search-based software engineering for software product lines. Information and Software Technology, 61, 33–51, 2015.

[S96] Lucas, F. J., Molina, F., and Toval, A. A systematic review of UML model consistency management. Information and Software Technology, 51(12), 1631–1645, 2009.

[S97] Machado, I. d., McGregor, J. D., Cavalcanti, Y. C., and de Almeida, E. S. On strategies for testing software product lines: A systematic literature review. Information and Software Technology, 56(10), 1183–1199, 2014.

[S98] Magdaleno, A. M., Werner, C. M., and Araujo, R. M. Reconciling software development models: A quasi-systematic review. Journal of Systems and Software, 85(2), 351–369, 2012.

[S99] Mahdavi-Hezavehi, S., Galster, M., and Avgeriou, P. Variability in quality attributes of service-based software systems: A systematic literature review. Information and Software Technology, 55(2), 320–343, 2013.

[S100] Maplesden, D., Tempero, E., Hosking, J., and Grundy, J. Performance Analysis for Object-Oriented Software: A Systematic Mapping. IEEE Transactions on Software Engineering, PP(99), 1–1, 2015.

[S101] Martin, W., Sarro, F., Jia, Y., Zhang, Y., and Harman, M. A Survey of App Store Analysis for Software Engineering. IEEE Transactions on Software Engineering, 1–1, 2016.

[S102] Martins, L. E., and Gorschek, T. Requirements engineering for safety-critical systems: A systematic literature review. Information and Software Technology, 75, 71–89, 2016.

[S103] Mehmood, A., and Jawawi, D. N. Aspect-oriented model-driven code generation: A systematic mapping study. Information and Software Technology, 55(2), 395–411, 2013.

[S104] Misbhauddin, M., and Alshayeb, M. UML model refactoring: a systematic literature review. Empirical Software Engineering, 20(1), 206–251, 2013.

[S105] Misbhauddin, M., and Alshayeb, M. UML model refactoring: a systematic literature review. Empirical Software Engineering, 20(1), 206–251, 2015.

[S106] Mohabbati, B., Asadi, M., Gaševic, D., Hatala, M., and Müller, H. A. Combining service-orientation and software product line engineering: A systematic mapping study. Information and Software Technology, 55(11), 1845–1859, 2013.

[S107] Mohagheghi, P., Dehlen, V., and Neple, T. Definitions and approaches to model quality in model-based software development – A review of literature. Information and Software Technology, 51(12), 1646–1669, 2009.

[S108] Mohamad Yusop, N. S., Grundy, J., and Vasa, R. Reporting Usability Defects: A Systematic Literature Review. IEEE Transactions on Software Engineering, 1–1, 2016.

[S109] Molleri, J. S., p, K., and Mendes, E. Survey Guidelines in Software Engineering: An Annotated Review. Proceedings of the 10th ACMIEEE International Symposium on Empirical Software Engineering and Measurement, 58:1–58:6, New York, NY, USA: ACM, 2016.

[S110] Montalvillo, L., and Di¬az, O. Requirement-driven evolution in software product lines: A systematic mapping study. Journal of Systems and Software, 122, 110–143, 2016.

[S111] Munir, H., Moayyed, M., and Petersen, K. Considering rigor and relevance when evaluating test driven development: A systematic review. Information and Software Technology, 56(4), 375–394, 2014.

[S112] Nair, S., de la Vara, J. L., Sabetzadeh, M., and Briand, L. An extended systematic literature review on provision of evidence for safety certification. Information and Software Technology, 56(7), 689–717, 2014.

[S113] Neiva, F. W., David, J. M., Braga, R., and Campos, F. Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature. Information and Software Technology, 72, 137–150, 2016.

[S114] Nguyen, P. H., Kramer, M., Klein, J., and Traon, Y. L. An extensive systematic review on the Model-Driven Development of secure systems. Information and Software Technology, 68, 62–81, 2015.

[S115] Nguyen-Duc, A., Cruzes, D. S., and Conradi, R. The impact of global dispersion on coordination, team performance and software quality – A systematic literature review. Information and Software Technology, 57, 277–294, 2015.

[S116] Novais, R. L., Torres, A., Mendes, T. S., MendonÃ§a, M., and Zazworka, N. Software evolution visualization: A systematic mapping study. Information and Software Technology, 55(11), 1860–1883, 2013.

[S117] Paternoster, N., Giardino, C., Unterkalmsteiner, M., Gorschek, T., and Abrahamsson, P. Software development in startup companies: A systematic mapping study. Information and Software Technology, 56(10), 1200–1218, 2014.

[S118] Penzenstadler, B., Bauer, V., Calero, C., & Franch, X. (2012, May). Sustainability in software engineering: A systematic literature review. In 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012), (pp. 32–41). IET.

[S119] Penzenstadler, B., Raturi, A., Richardson, D., Calero, C., Femmer, H., & Franch, X. (2014, May). Systematic mapping study on software engineering for sustainability (SE4S). In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (p. 14). ACM.

[S120] Petersen, K. Measuring and predicting software productivity: A systematic map and review. Information and Software Technology, 53(4), 317–343, 2011.

[S121] Pretorius, R., and Budgen, D. A Mapping Study on Empirical Evidence Related to the Models and Forms Used in the Uml. Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 342–344, New York, NY, USA: ACM, 2008.

[S122] Qureshi, N., Usman, M., & Ikram, N. (2013, April). Evidence in software architecture, a systematic literature review. In Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering (pp. 97–106). ACM.

[S123] Rabiser, R., Grünbacher, P., and Dhungana, D. Requirements for product derivation support: Results from a systematic literature review and an expert survey. Information and Software Technology, 52(3), 324–346, 2010.

[S124] Radjenovic, D., Heriv cko, M., Torkar, R., and Živkoviv c, A. Software fault prediction metrics: A systematic literature review. Information and Software Technology, 55(8), 1397–1418, 2013.

[S125] Rafique, Y., & Mišić, V. B. (2013). The effects of test-driven development on external quality and productivity: A meta-analysis. IEEE Transactions on Software Engineering, 39(6), 835–856.

[S126] Rashid, M., Anwar, M. W., and Khan, A. M. Toward the tools selection in model based system engineering for embedded systems*β*€” A systematic literature review. Journal of Systems and Software, 106, 150–163, 2015.

[S127] Riaz, M., Maintainability prediction of relational database-driven applications: A systematic review. In 16th International Conference on the Evaluation & Assessment in Software Engineering (EASE 2012), (pp. 263–272). IET.

[S128] Riaz, M., Breaux, T., and Williams, L. How have we evaluated software pattern application? A systematic mapping study of research design practices. Information and Software Technology, 65, 14–38, 2015.

[S129] Rickckevivcs, K., and Torkar, R. Equality in cumulative voting: A systematic review with an improvement proposal. Information and Software Technology, 55(2), 267–287, 2013.

[S130] Salleh, N., Mendes, E., and Grundy, J. Empirical Studies of Pair Programming for CSSE Teaching in Higher Education: A Systematic Literature Review. IEEE Transactions on Software Engineering, 37(4), 509–525, 2011.

[S131] Santos Rocha, R. d., and Fantinato, M. The use of software product lines for business process management: A systematic literature review. Information and Software Technology, 55(8), 1355–1373, 2013.

[S132] Santos, A. R., de Oliveira, R. P., & de Almeida, E. S. (2015, April). Strategies for consistency checking on software product lines: a mapping study. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (p. 5). ACM.

[S133] Santos, P. S., and Travassos, G. H. Action Research Use in Software Engineering: An Initial Survey. Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, 414–417, Washington, DC, USA: IEEE Computer Society, 2009.

[S134] Santos, R. E., da Silva, F. Q., & de Magalhães, C. V., Benefits and limitations of job rotation in software organizations: a systematic literature review. In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering (p. 16). ACM.

[S135] Saraiva, J., Barreiros, E., Almeida, A., Lima, F., Alencar, A., Lima, G, & Castor, F. (2012, May). Aspect-oriented software maintenance metrics: A systematic mapping study. In Proceedings of the 16th International Conference on the Evaluation & Assessment in Software Engineering (EASE 2012), (pp. 253–262). IET.

[S136] Senapathi, M. and Srinivasan, A., "Sustained agile usage: a systematic literature review. In Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering (EASE ' 13). ACM, New York, NY, USA, 119–124.

[S137] Sepulveda, S., Cravero, A., and Cachero, C. Requirements modeling languages for software product lines: A systematic literature review. Information and Software Technology, 69, 16–36, 2016.

[S138] Shahin, M., Liang, P., and Babar, M. A. A systematic review of software architecture visualization techniques. Journal of Systems and Software, 94, 161–185, 2014.

[S139] Shahrokni, A., and Feldt, R. A systematic review of software robustness. Information and Software Technology, 55(1), 1–17, 2013.

[S140] Sharafi, Z., Soh, Z., and GuΓ©hΓ©neuc, Y.-G. A systematic literature review on the usage of eye-tracking in software engineering. Information and Software Technology, 67, 79–107, 2015.

[S141] Shippey, T., Bowes, D., Chrisianson, B., & Hall, T. “A mapping study of software code cloning” . In 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012).

[S142] Siegmund, J., and Schumann, J. Confounding parameters on program comprehension: a literature survey. Empirical Software Engineering, 20(4), 1159–1192, 2015.

[S143] Silva, F. Q., Suassuna, M., França, A. C., Grubb, A. M., Gouveia, T. B., Monteiro, C. V. Replication of empirical studies in software engineering research: a systematic mapping study. Empirical Software Engineering, 19(3), 501–557, 2012.

[S144] Smite, D., Wohlin, C., Gorschek, T., and Feldt, R. Empirical evidence in global software engineering: a systematic review. Empirical Software Engineering, 15(1), 91–118, 2009.

[S145] Soomro, A. B., Salleh, N., Mendes, E., Grundy, J., Burch, G., and Nordin, A. The effect of software engineers personality traits on team climate and performance: A Systematic Literature Review. Information and Software Technology, 73, 52–65, 2016.

[S146] Souza Neto, P. A., Vargas-Solar, G., da Costa, U. S., and Musicante, M. A. Designing service-based applications in the presence of non-functional properties: A mapping study. Information and Software Technology, 69, 84–105, 2016.

[S147] Staples, M., and Niazi, M. Systematic review of organizational motivations for adopting CMM-based SPI. Information and Software Technology, 50(7–8), 605–620, 2008.

[S148] Steinmacher, I., Graciotto Silva, M. A., Gerosa, M. A., and Redmiles, D. F. A systematic literature review on the barriers faced by newcomers to open source software projects. Information and Software Technology, 59, 67–85, 2015.

[S149] Stevanetic, S., & Zdun, U. (2015, April). Software metrics for measuring the understandability of architectural structures: a systematic mapping study. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (p. 21). ACM.

[S150] Stol, K.-J., Ralph, P., and Fitzgerald, B. Grounded Theory in Software Engineering Research: A Critical Review and Guidelines. Proceedings of the 38th International Conference on Software Engineering, 120–131, Austin, Texas: ACM, 2016.

[S151] Sulaman, S. M., Weyns, K., & Höst, M. (2013, April). “A review of research on risk analysis methods for IT systems” . In Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering (pp. 86–96). ACM.

[S152] Tahir, A., Tosi, D., and Morasca, S. A systematic review on the functional testing of semantic web services. Journal of Systems and Software, 86(11), 2877–2889, 2013.

[S153] Tahir, T., Rasool, G., and Gencel, C. A systematic literature review on software measurement programs. Information and Software Technology, 73, 101–121, 2016.

[S154] Tarhan, A., Turetken, O., and Reijers, H. A. Business process maturity models: A systematic literature review. Information and Software Technology, 75, 122–134, 2016.

[S155] Tiwari, S., and Gupta, A. A systematic literature review of use case specifications research. Information and Software Technology, 67, 128–158, 2015.

[S156] Tofan, D., Galster, M., Avgeriou, P., and Schuitema, W. Past and future of software architectural decisions – A systematic mapping study. Information and Software Technology, 56(8), 850–872, 2014.

[S157] Tosi, D., and Morasca, S. Supporting the semi-automatic semantic annotation of web services: A systematic literature review. Information and Software Technology, 61, 16–32, 2015.

[S158] Turner, M., Kitchenham, B., Brereton, P., Charters, S., and Budgen, D. Does the technology acceptance model predict actual use? A systematic literature review. Information and Software Technology, 52(5), 463–479, 2010.

[S159] Unterkalmsteiner, M., Gorschek, T., Islam, A., Cheng, C. K., Permadi, R., and Feldt, R. Evaluation and Measurement of Software Process Improvement—A Systematic Literature Review. IEEE Transactions on Software Engineering, 38(2), 398–424, 2012.

[S160] Wen, J., Li, S., Lin, Z., Hu, Y., and Huang, C. Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology, 54(1), 41–59, 2012.

[S161] Wnuk, K., & Kollu, R. K. (2016, June). A systematic mapping study on requirements scoping. In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering (p. 32). ACM.

[S162] Yang, C., Liang, P., and Avgeriou, P. A systematic mapping study on the combination of software architecture and agile development. Journal of Systems and Software, 111, 157–184, 2016.

[S163] Zarour, M., Abran, A., Desharnais, J.-M., and Alarifi, A. An investigation into the best practices for the successful design and implementation of lightweight software process assessment methods: A systematic literature review. Journal of Systems and Software, 101, 180–192, 2015.

[S164] Zein, S., Salleh, N., and Grundy, J. A systematic mapping study of mobile application testing techniques. Journal of Systems and Software, 117, 334–356, 2016.

[S165] Zhi, J., Garousi-Yusifov glu, V., Sun, B., Garousi, G., Shahnewaz, S., and Ruhe, G. Cost, benefits and quality of software development documentation: A systematic mapping. Journal of Systems and Software, 99, 175–198, 2015.

Appendix C. Venues Selection Process

Name	cr.1	cr.2	cr.3	cr.4	Included
IEEE Transactions on Software Engineering	A	yes	yes	183	yes
International Conference on Software Engineering	A	yes	yes	118	yes
IEEE Software	B	yes	yes	108	yes
Software: Practice and Experience	A	yes	yes	80	yes
ACM Transactions on Software Engineering and Methodology	A	yes	yes	69	yes
Journal of Systems and Software	A	yes	yes	61	yes
Automated Software Engineering	A	yes	yes	53	yes
Information and Software Technology	B	yes	yes	46	yes
European Software *Engineering* Conference and the ACM SIGSOFT International Symposium on the Foundations of Software Engineering	A	yes	yes	44	yes
Automated Software Engineering Conference	A	yes	yes	44	yes
Empirical Software Engineering	A	yes	yes	36	yes
International Symposium on Empirical Software Engineering and Measurement	A	yes	yes	21	yes
ACM Computing Surveys	A	no			no
ACM Transactions on Architecture and Code Optimization	A	yes	no		no
ACM Transactions on Computer Systems	A	no			no
ACM Transactions on Design Automation of Electronic Systems	A	no			no
ACM Transactions on Embedded Computing Systems	A	no			no
ACM Transactions on Information and System Security	A	yes	no		no
ACM Transactions on Multimedia Computing Communications and Applications	B	yes	no		no
ACM Transactions on Programming Languages and Systems	A	yes	no		no
Acta Informatica	A	yes	yes	N/A	no
Computer Standards and Interfaces	B	no			no
Computers and Electrical Engineering	B	no			no
Computers and Security	B	yes	no		no
Computers in Industry	B	no			no
IBM Journal of Research and Development	A	no			no
IBM Systems Journal	A	no			no
IEEE Transactions on Computers	A	no			no
IEEE Transactions on Dependable and Secure Computing	A	no			no
IEEE Transactions on Multimedia	A	yes	no		no
IEEE Transactions on Reliability	A	yes	no		no
IET Computers and Digital Techniques	B	no			no
Industrial Management + Data Systems	B	no			no
Innovations in Teaching and Learning in Information and Computer Sciences	B	no			no
International Journal of Agent Oriented Software Engineering	B	yes	no		no
International Journal on Software Tools for Technology Transfer	B	yes	no		no
Journal of Computer Security	B	no			no
Journal of Functional and Logic Programming	B	yes	no		no
Journal of Object Technology	B	yes	no		no

(*continued on next page*)

(*continued*)

Name	cr.1	cr.2	cr.3	cr.4	Included
Journal of Software	B	yes	yes	N/A	no
Journal of Software Maintenance and Evolution: research and practice	B	yes	no		no
Journal of Systems Architecture	B	yes	no		no
Journal of Visual Languages and Computing	A	yes	no		no
Multimedia Systems	B	yes	no		no
Multimedia Tools and Applications	B	yes	no		no
Requirements Engineering	B	yes	no		no
Science of Computer Programming	A	yes	no		no
Software and System Modelling	B	yes	no		no
Software Testing, Verification and Reliability	B	yes	no		no
Text Technology: the journal of computer text processing	B	no			no
Theory and Practice of Logic Programming	A	yes	no		no
ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication	A	no			no
ACM Conference on Computer and Communications Security	A	no			no
ACM Conference on Object Oriented Programming Systems Languages and Applications	A	yes	no		no
ACM International Symposium on Computer Architecture	A	yes	no		no
ACM Multimedia	A	no			no
ACM SIGOPS Symposium on Operating Systems Principles	A	no	no		no
ACM/IFIP/USENIX International Middleware Conference	A	no			no
ACM-SIGACT Symposium on Principles of Programming Languages	A	yes	no		no
ACM-SIGPLAN Conference on Programming Language Design and Implementation	A	yes	no		no
Annual Computer Security Applications Conference	A	yes	no		no
Architectural Support for Programming Languages and Operating Systems	A	yes	no		no
Aspect-Oriented Software Development	A	yes	no		no
Conference on the Quality of Software Architectures	A	yes	no		no
European Conference on Object-Oriented Programming	A	yes	no		no
European Symposium on Programming	A	yes	no		no
European Symposium On Research In Computer Security	A	yes	no		no
Eurosys Conference	A	yes	no		no
IEEE Computational Systems Bioinformatics Conference	A	no			no
IEEE Computer Security Foundations Symposium	A	yes	no		no
IEEE International Conference on Software Maintenance	A	yes	no		no
IEEE International Requirements Engineering Conference	A	yes	no		no
IEEE/IFIP International Conference on Dependable Systems	A	yes	no		no
IEEE/IFIP International Symposium on Trusted Computing and Communications	A	no			no
IEEE/IFIP Working Conference on Software Architecture	A	yes	no		no
IFIP Joint International Conference on Formal Description Techniques and Protocol Specification, Testing, And Verification	A	yes	no		no
Intelligent Systems in Molecular Biology	A	no			no
International Conference on Compiler Construction	A	yes	no		no
International Conference on Coordination Models and Languages	A	yes	no		no
International Conference on Evaluation and Assessment in Software Engineering	A	yes	yes	N/A	no
International Conference on Functional Programming	A	yes	no		no
International Conference on Principles and Practice of Constraint Programming	A	yes	no		no
International Conference on Reliable Software Technologies	A	yes	no		no
International Conference on Software Process	A	yes	no		no
International Conference on Security and Privacy for Communication Networks	A	no			no
International Conference on Software Reuse	A	yes	no		no
International Conference on Virtual Execution Environments	A	no			no
International Symposium Component-Based Software Engineering	A	yes	no		no
International Symposium on Automated Technology for Verification and Analysis	A	yes	no		no
International Symposium on Code Generation and Optimization	A	yes	no		no
International Symposium on High Performance Computer Architecture	A	yes	no		no
International Symposium on Memory Management	A	yes	no		no
International Symposium on Software Reliability Engineering	A	yes	no		no
International Symposium on Software Testing and Analysis	A	yes	no		no
Tools and Algorithms for Construction and Analysis of Systems	A	yes	no		no
Usenix Network and Distributed System Security Symposium	A	yes	no		no
Usenix Security Symposium	A	yes	no		no
Usenix Symposium on Operating Systems Design and Implementation	A	no			no
USENIX Workshop on Hot Topics in Operating Systems	A	no			no

References

[1] S.A. Avellar, J. Thomas, R. Kleinman, E. Sama-Miller, Woodruff S.E, R. Coughlin, T.P.R Westbrook, External validity: the next step for systematic reviews? *Eval. Rev.* 41 (4) (2017) 283–325.

[2] M. Bano, D. Zowghi, N. Ikram, Systematic reviews in requirements engineering: a tertiary study", Empirical Requirements Engineering (EmpiRE), in: 2014 IEEE Fourth International Workshop on, August 2014, pp. 9–16.

[3] V.R. Basili, R.W. Selby, Paradigms for experimentation and empirical studies in software engineering, Reliab. Eng. Syst. Saf. 32 (1–2) (1991) 171–191.

[4] S. Biffl, M. Kalinowski, F. Ekaputra, A.A. Neto, T. Conte, D. Winkler, Towards a semantic knowledge base on threats to validity and control actions in controlled experiments, 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), ACM, 2014.

[5] D. Budgen, P. Brereton, S. Drummond, N. Williams, Reporting systematic reviews: some lessons from a tertiary study, *Inf. Softw. Technol.* 95 (2018) 62–74.

[6] K.Y. Cai, David Card, An analysis of research topics in software engineering – 2006, J. Syst. Softw. 81 (6) (June 2008) 1051–1058.

[7] Centre for Reviews and Dissemination, U. O. Y., The Database of Abstracts of Reviews of Effects (DARE), Effect. Matters 6 (2) (2002) 1–4.

[8] T.D. Cook, D.T. Campbell, Quasi-experimentation: design & analysis issues for field settings, *Boston* (1979).

[9] D.S. Cruzes, T. Dybå, Research synthesis in software engineering: a tertiary study, Inf. Softw. Technol. 53 (5) (May 2011) 440–455.

[10] S.H. Downs, N. Black, The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions, Journal of Epidemiology & Community Health 52 (6) (1998) 377–384.

[11] K. Dwan, D.G. Altman, J.A. Arnaiz, J. Bloom, A.W. Chan, E. Cronin, E. Decullier, PJ. Easterbrook, E. Von Elm, C. Gamble, D. Ghersi, JP. Ioannidis, J. Simes, PR. Williamson, D. Ghersi, "Systematic review of the empirical evidence of study publication bias and outcome reporting bias, PLoS One 3 (8) (2008) e3081.

[12] R. Feldt, A. Magazinius, Validity Threats in Empirical Software Engineering Research - An Initial Survey, in: Proceedings of the 22nd Int. Conf. on Software Engineering and Knowledge Engineering (SEKE), Redwood City, California, July 2010, pp. 374–379.

[13] M. Galster, D. Weyns, D. Tofan, B. Michalik, P. Avgeriou, Variability in software systems - A systematic literature review, IEEE Trans. Softw. Eng. 40 (3) (2014) 282–306.

[14] D.J. Greenwood, M. Levin, Introduction to Action research: Social Research For Social Change, 2nd ed, SAGE, Thousand Oaks, Calif., 2007.

[15] S. Imtiaz, M. Bano, N. Ikram, 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, 2013.

[16] A.R. Jadad, R.A. Moore, D. Carroll, C. Jenkinson, D.J.M. Reynolds, D.J. Gavaghan, H.J. McQuay, Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clin. Trials 17 (1) (1999) 1–12.

[17] A. Jedlitschka, D. Pfahl, Reporting guidelines for controlled experiments in software engineering, in: Proceedings of the International Symposium on Empirical Software Engineering, IEEE, 2005, pp. 10–17.

[18] B.A. Kitchenham, S. Charters, Guidelines For Performing Systematic Literature Reviews in Software Engineering, School of Computer Science and Mathematics, Keele University., 2007 *Technical Report EBSE-2007-01*.

[19] B.A. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, S. Linkman, Systematic literature reviews in software engineering – A tertiary study, Inf. Softw. Technol. 52 (8) (August 2010) 792–805.

[20] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – a systematic literature review, Inf. Softw. Technol. 51 (1) (January 2009) 7–15.

[21] B. Kitchenham, D. Budgen, P. Brereton, The value of mapping studies – A participant-observer case study, in: Evaluation and Assessment in Software Engineering (EASE '10), UK, British Computer Society Swinton, 2010, pp. 1–9.

[22] B.A. Kitchenham, T. Dybå, M. Jørgensen, Evidence-based software engineering, in: Proceedings of the 26th International Conference on Software Engineering (ICSE '04), IEEE, May , 2004, pp. 273–281.

[23] B. Kitchenham et al. "The impact of limited search procedures for systematic literature reviews: a participant-observer case study", Proc. Third Int'l Symp. Empirical Software Eng. and Measurement, pp. 336–345

[24] B. Kitchenham, P. Bereton, A systematic review of systematic review process research in software engineering, *Inf. Softw. Technol.* 55 (12) (2013) 2049–2075.

[25] B. Kitchenham, D. Budgen, P. Brereton, Evidence Base Software Engineering and Systematic Reviews, Chapman and Hall/CRC, 2015.

[26] T.C. Lethbridge, S.E. Sim, J. Singer, Studying software engineers: data collection techniques for software field studies, Empirical Softw. Eng. 10 (3) (Jul. 2005) 311–341.

[27] Y.S. Lincoln, E.G. Guba, Naturalistic Inquiry, Sage, Beverly Hills, Calif., 1985.

[28] A.B. Marques, R. Rodrigues, T. Conte, Systematic literature reviews in distributed software development: a tertiary study, Global Software Engineering (ICGSE '12), IEEE, August 2012 27-30.

[29] J.A. Maxwell, Understanding and validity in qualitative research, Harvard educational review 62 (3) (1992) 279–301.

[30] R. Nickerson, J. Muntermann, U. Varshney, H. Isaac, Taxonomy development in information systems: developing a taxonomy of mobile applications, European Conference in Information Systems (2009).

[31] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L.A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement, Systematic Reviews 54 (1) (2015).

[32] D.E. Perry, A.A. Porter, L.G. Votta, Empirical studies of software engineering: a roadmap, in: Proceedings of the conference on the future of Software engineering, ACM, 2000, pp. 345–355.

[33] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, in: In the proceedings of Evaluation and Assessment in Software Engineering, 8, EASE, 2008, pp. 68–77.

[34] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: an update, Information and Software Technology 64 (August 2015) 1–18.

[35] K. Petersen, C. Gencel., Worldviews, research methods, and their relationship to validity in empirical software engineering research, in: *Proceedings of the Eighth International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, 2013, Oct. 2013, pp. 81–89.

[36] S.L. Pfleeger, B.A. Kitchenham, Principles of survey research: part 1: turning lemons into lemonade, *SIGSOFT Softw. Eng. Notes 26*, 6 November 2001, 2001.

[37] H. Rothstein, Publication bias as a threat to the validity of meta-analytic results, J. Exp. Criminol. (2008) 61–81 4.1.

[38] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Emp. Softw. Eng. 14 (2) (December 2009) 131–164.

[39] B.J. Shea, J.M. Grimshaw, G.A. Wells, M. Boers, N. Andersson, C. Hamel, A.C. Porter, P. Tugwell, D. Moher, L.M. Bouter, Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews, BMC Med. Res. Methodol. 7 (1) (2007) 10.

[40] J. Siegmund, N. Siegmund, S. Apel, Views on internal and external validity in empirical software engineering, in: 37th IEEE International Conference on Software Engineering (ICSE), Florence, Italy, ACM, 2015, pp. 9–19.

[41] F.Q.B. da Silva, A.L.M. Santos, S. Soares, A.C.C. França, C.V.F. Monteiro, F.F. Maciel, Six years of systematic literature reviews in software engineering: an updated tertiary study, Inf. Softw. Technol. 53 (9) (2011) 899–913.

[42] D. Sjoberg, T. Dyba, M. Jorgensen, The future of empirical methods in software engineering research, in: Proceedings of 2007 Future of Software Engineering, IEEE Computer Society, 2007, pp. 358–378.

[43] T.A. Slocum, R. Detrich, T.D. Spencer, "Evaluating the validity of systematic reviews to indentify empirically supported treatments, Educ. Treat. Child. 35 (2) (2012) 201–233.

[44] W.F. Tichy, F. Padberg, Empirical methods in software engineering research, ICSE Companion (2007) 163–164.

[45] A.P. Verhagen, H.C. de Vet, R.A. de Bie, A.G. Kessels, M. Boers, L.M. Bouter, P.G. Knipschild, The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus, J. Clin. Epidemiol. 51 (12) (1998) 1235–1241.

[46] J.M. Verner, O.P. Brereton, B.A. Kitchenham, M. Turner, Systematic literature reviews in global software development: a tertiary study, Evaluation and Assessment in Software Engineering (EASE '12), IET, 14-15 May 2012.

[47] C. Wohlin, M. Host, P. Runeson, M. Ohlsson, B. Regnell, and A. Wesslen, "Experimentation in software engineering: an introduction", Kluwer Academic Publishers, 2000

[48] C. Wohlin, P. Runeson, P. Anselmo da Mota Silveira Neto, E. Engström, I. do Carmo Machado, E. Santana de Almeida, On the reliability of mapping studies in software engineering, J. Syst. Softw. 86 (10) (2013) 2594–2610.

[49] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14), New York, NY, USA, ACM, 2014 Article 38, 10 pages.

[50] W.E. Wong, T.H. Tse, R.L. Glass, V.R. Basili, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (2003–2007 and 2004–2008), J. Syst. Softw. 84 (1) (January 2011) 162–168.

[51] R.K. Yin, Case Study research: Design and Methods, 4th ed, SAGE, London, 2009.

[52] S. Zaza, L.K. Wright-De Agüero, P.A. Briss, B.I. Truman, D.P. Hopkins, M.H. Hennessy, D.M. Sosin, L. Anderson, V.G. Carande-Kulis, S.M. Teutsch, M. Pappaioanou, Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services, Am. J. Prevent. Med. 18 (1) (2000) 44–74.

[53] X. Zhou, Y. Jin, H. Zhang, S. Li, X. Huang, A map of threats to validity of systematic literature reviews in software engineering, in: 23rd Asia-Pacific Software Engineering Conference (APSEC), Hamilton, 2016, pp. 153–160.