Guidelines for Managing Threats to Validity of Secondary Studies in Software Engineering[1]

Apostolos Ampatzoglou[1], Stamatia Bibi[2], Paris Avgeriou[3], Alexander Chatzigeorgiou[1]

[1] Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece

[2] Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece

[3] Department of Mathematics and Computer Science, University of Groningen, the Netherlands

Abstract

Secondary studies review and compile data retrieved from primary studies and are vulnerable to factors that threaten their validity as any other research method. Considering that secondary studies are often used to support the evidence-based paradigm, it is crucial to properly manage their threats, i.e., identify, categorize, mitigate, and report them. In this chapter, we build upon the outcomes of a systematic review of secondary studies in software engineering, which has identified: (a) the most common threats to validity and corresponding mitigation actions; and (b) the categories in which threats to validity can be classified, so as to guide the authors of future secondary studies in managing the threats to validity of their work. To achieve this goal, we describe: (a) a classification schema for reporting threats to validity and possible mitigation actions; and (b) 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

Over the past decade, due to the rise of the Evidence-Based Software Engineering (EBSE) Paradigm [34], there has been a proliferation of secondary studies. In this chapter we focus on two types of secondary studies:

- ***Systematic Literature Reviews (SLRs),*** which constitute the core tool of the evidence-based paradigm and originated in clinical medicine. SLRs aim at gathering data from previously published studies, objectively and without bias, for the purpose of synthesizing existing evidence and answering research questions. Re-

[1] Based on Ampatzoglou et al. [8]: Identifying, categorizing and mitigating threats to validity in software engineering secondary studies, *Information and Software Technology*, Elsevier, 106 (2), pp. 201–230, February 2019.

search synthesis is a collective term for a family of methods for summarizing, integrating and, when possible, combining the findings of different studies. 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 empirical findings from individual studies are limited in the extent to which they may be generalized [31].

- ***Systematic Mapping Studies (SMS)***, which use the same basic methodology as SLRs but aim to provide a more coarse-grained overview of the research that has been performed on a topic rather than answering questions about the relative merits of competing technologies. In a SMS, published results are usually mapped onto a classification schema and visualized focusing on frequencies of publications for sub-topics within the schema [45].

EBSE research relies substantially on systematic and rigorous guidelines on how to conduct, and report empirical results—e.g., experiments [56], SLRs [31], SMSs [45], surveys [46], case studies [47]. These guidelines emphasize, among others, the importance of managing (identifying, managing and reporting) relevant *threats to validity*, i.e., possible aspects of the research design that in some way compromise the credibility of results. However, we currently lack guidelines on how to manage threats to validity in secondary studies. In this chapter, we build upon the results of a tertiary study (i.e., an SLR on secondary studies in software engineering) [8], namely a classification schema for threats to validity and corresponding mitigation actions, combined with a checklist to be used while conducting/evaluating secondary studies.

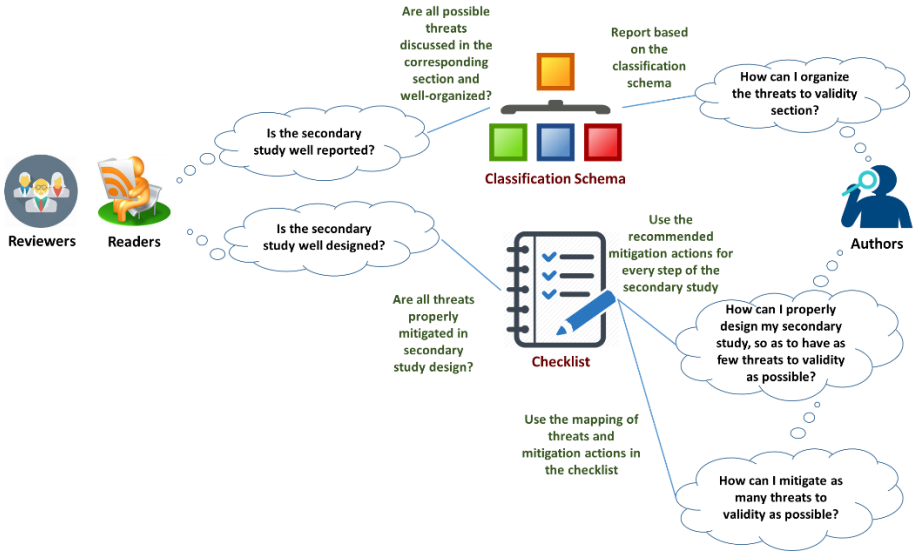

Fig. 1. Usage Scenarios of Threats to Validity Guidelines

The classification schema and the checklist can assist different stakeholders with various activities as illustrated in Figure 1. First, we expect that a critical appraisal

of secondary studies can be performed by readers and reviewers, by consulting the checklist to identify possible threats in the study design, and confirm that they have been properly mitigated. Also, the reporting of the studies can be evaluated, both in terms of threats to validity and their mitigation, as well as in terms of categorization. Second, authors of secondary studies can be guided on how to setup their study design, so as to avoid or mitigate validity threats, while planning, conducting and reporting secondary studies.

Section 2 presents the basis of this chapter, i.e., the classification schema and the validity checklist, proposed by Ampatzoglou et al. [8]. In Section 3, we present the first usage scenario, which exemplifies how the classification schema and the checklist can be used by authors of secondary studies, whereas Section 4 discusses the usage scenario for reviewers and readers of secondary studies. Finally, Section 5 discusses further readings, and Section 6 concludes the chapter.

2. Classification Schema and Validity Checklist

Identifying, classifying and mitigating threats to the validity of results obtained through secondary studies are important to increase our confidence on the conclusions drawn from these results. Despite the fact that the percentage of secondary studies reporting threats to validity has been continuously increasing, considerable confusion still exists in terms of terminology, mitigation strategies, and classification [8] often leading to erroneous classification of threats. For instance, in many secondary studies any bias that might be introduced during study selection, is wrongly classified (by the authors of secondary studies) under *internal* validity almost as often as under *reliability,* pointing to inconsistencies in the classification of threats [8]. 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, thus posing a threat to 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 posing reliability threats. While one can argue about the correctness of both classifications, multi-label classification can be confusing and does not allow for a uniform comparison of the threats. Therefore, next we present a classification schema for threats to validity and their mitigation actions, tailored for secondary studies.

2.1 Classification Schema

The classification schema consists of three levels: the first one depicting threat categories, the second, threats per se, whereas the third one, mitigation actions. To derive the threat categories (first level of the schema) and to facilitate the classification of any given threat, we use the planning phases of the secondary studies (i.e., search process, study filtering, data extraction and analysis—see Figure 2). These are easily

identifiable steps in a secondary study, in contrast to using the aspects of validity that are threatened (e.g., internal / external / construct validity, etc.). Moreover, we have added an additional category (i.e., a horizontal one) that corresponds to threats that cover the lifecycle of the secondary study:

- ***Study Selection Validity***. This category involves threats that can be identified in the first two phases of secondary studies (i.e., search process and study filtering phase). Issues classified in this category threaten the validity of searching and including primary studies in the examined set. This involves threats like the *selection of digital libraries*, *search string construction*, 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 *small sample size*, *lack of statistical analysis,* etc.
- ***Research Validity***. Threats that can be identified in all phases and concern the overall research design are classified into this category. Examples of threats in this category are: *generalizability, coverage of research questions*, etc.

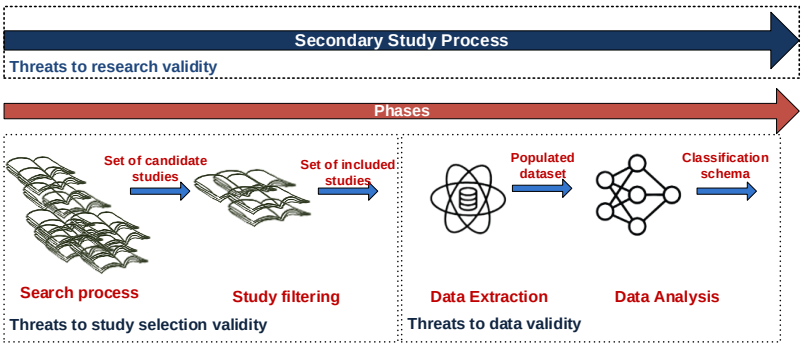

Fig. 2. Secondary Studies Phases and Corresponding Threats

Although the majority of the names for threats to validity and mitigation actions can be considered self-explanatory, more details are provided in Section 3. We note that due to space limitations, only the most frequent mitigation actions for every threat are presented in Fig. 3a-c. The full list of mitigation actions is available online, in the accompanying technical report of Ampatzoglou et al. [8]. The three categories of validity threats along with the proposed mitigation actions are shown in Figures 3a to 3c respectively. Blue cells represent threats to validity and red cells to mitigation actions. Groups of validity threats are depicted as adjacent blue cells. When a number of mitigation actions can be used for a threat or a group of threats, they are also depicted as adjacent red cells.

The ***study selection validity*** category involves 7 specific threats (see Fig. 3a). Five threats to validity can be grouped in a more generic one, i.e., *Adequacy of initial relevant publication identification* (TV_1), whereas the rest are ungrouped. From the threats of this category, some are mutually exclusive, whereas others may coexist.

For example, if *selection of digital libraries* is performed, the threat *selection of publication venues* ($TV_{1.3}$) is excluded since, normally only one of the two search strategies (digital libraries or venues) 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 ($TV_{1.1}$) exists both when digital libraries or specific publication venues are selected. After the initial set of publications is derived, other aspects threaten the validity of the study: *were there enough journals and conferences for the authors to search* (TV_2), *what languages have the authors explored* (TV_3), *were all papers accessible by the authors* (TV_4), *how have the authors handled the duplicate articles* (TV_5) or *the grey literature* (TV_6), and is the *selection of inclusion/exclusion criteria accurate*? (TV_7).

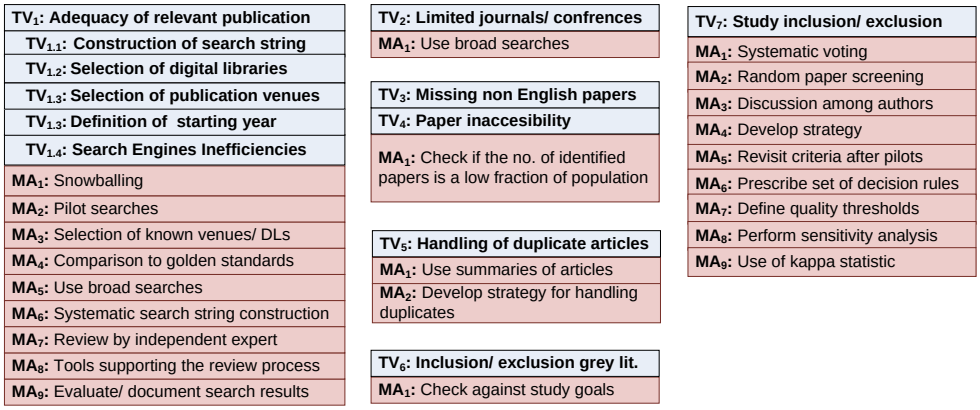

Fig. 3a. Study Selection Validity Threats and Mitigation Actions

The ***data validity*** category includes 9 specific threats (see Fig. 3b), that are organized into three groups and five ungrouped threats to validity. One group includes any kind of bias that can be introduced while collecting data, namely: *data extraction bias* (TV_{13}), *data extraction inaccuracies, quality assessment subjectivity, unverified data extraction*, and *misclassification of primary studies* (mostly relevant for mapping studies). Another group includes limitations of the dataset (TV_8) that are due to the nature of the subject and not due to researchers' bias (i.e., *small sample size* and *heterogeneous primary studies*). A third group represents threats that are relevant for mapping studies and have been posed by the use of *inadequate classification schemas* or *attributes frameworks* (TV_{15}). Furthermore, other aspects such as the *validity of primary studies* (TV_{12}), *the potential lack of relationships in the dataset* (TV_{11}), *the publication bias* (TV_{10}), and *the choice of extracted variables* (TV_9) 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* (TV_{16}) while interpreting the results and the *lack of statistical analysis* (TV_{14}).

TV_8: Small sample size
$TV_{8.1}$: Small sample size
$TV_{8.2}$: Primary stud. heterogeneity
MA_1: Draw conclusions based on trends
MA_2: Use broad searches

TV_9: Choices of variables to be extracted
MA_1:Discuss among authors

TV_{10}: Publication bias
MA_1: Snowballing
MA_2: Include grey literature
MA_3: Investigate manually other venues

TV_{11}: Lack of relationships
MA_1: Pilot data extraction

TV_{12}: Validity of primary studies
MA_1: Only quality venues
MA_2: Quality assessment
MA_3: Assess the validity using statistics

TV_{13}: Data extraction bias
$TV_{13.1}$: Data extraction bias
$TV_{13.2}$: Quality assessment subjectivity
$TV_{13.3}$: Data extraction inaccuracies
$TV_{13.4}$: Unverified data extraction
$TV_{13.5}$: Misclassification of studies
MA_1: Involve more than one researchers
MA_2: Use kappa statistic
MA_3: Pilot data extraction
MA_4: Use expert's opinion
MA_5: Random paper screening
MA_6: Perform keywording of abstracts

TV_{14}: Lack of statistical analysis
MA_1: Check for quantitative data

TV_{15}: Bias of Classification Schema
$TV_{15.1}$: Robustness of initial classification
TV_{15}: Construction of attribute framework
MA_1: Use existing schemas
MA_2: Continuous update

TV_{16}: Researcher bias
MA_1: Pilot data analysis
MA_2: Conduct reliability checks
MA_3: Use formal data synthesis
MA_4: Perform sensitivity analysis
MA_5: Use scientific quality of primary studies when drawing conclusions

Fig. 3b. Data Validity Threats and Mitigation Actions

Finally, the ***research validity*** category includes 6 specific threats (see Fig. 3c) that are forming two groups and include four ungrouped threats. The first group 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 (TV_{18}). Second, sometimes researchers *deviate from the established review process*. The second group involves threats to *generalizability* (TV_{22}). The individual threats that are mapped to this category are the *lack of comparable studies* (TV_{20}), the *coverage of research questions* (TV_{19}), and the *unfamiliarity of researchers with the application domain* (TV_{21}). Finally, *repeatability* (TV_{17}) 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.

TV_{17}: Repeatability
MA_1: Involve more than one researchers
MA_2: Make data available
MA_3: Develop protocol

TV_{18}: Research Method Bias
$TV_{18.1}$: Chosen research method
$TV_{18.2}$: Review process deviation
MA_1: Discuss among authors
MA_2: Develop protocol

TV_{19}: Coverage of research questions
MA_1: Brainstorming
MA_2: Motivate well research questions
MA_3: Consult target audience

TV_{20}: Lack of comparable studies
MA_1: Brainstorming

TV_{21}: Unfamiliar with research field
MA_1: Compare with related work

TV_{22}: Generalizability
$TV_{22.1}$: Generalizability
$TV_{22.2}$: Research not applicable to other domains/ organizations
MA_1: Use a broad search
MA_2: Compare with existing studies

Fig. 3c. Research Validity Threats and Mitigation Actions

Although we believe that the current classification schema improves the orthogonality among threat categories, there are still some "grey-zone" threats. Using the proposed classification schema, we address the problem of classifying a single threat to two categories: every threat is classified within one category, based on the phase of the study design, in which it was identified and the set of artifacts, whose validity is threatened. We identified five cases of threats that can be classified into more than one category:

- ***Quality Assessment Subjectivity***—In the context of secondary studies, the quality of a primary study can either be used as an inclusion criterion or as a variable that is collected during data extraction (when for example, the quality of the primary studies is part of the research questions) Thus, *Quality Assessment Subjectivity* can be classified under both Study Selection Validity and Data Validity, based on the role of the quality assessment. To ease the readability of this section, *Quality Assessment Subjectivity* is 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 or researchers involved in a particular scientific subdiscipline. At the same time, low *validity of primary studies* also 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.

2.2 Checklist for Threats to Validity Identification and Mitigation

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 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 each top-level 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. 3a (study selection validity). The mapping between questions and threats reported in Fig. 3 is one-to-one, by considering the groups discussed in Section 2.1.

Study Selection Validity

TV_1: Has your search process adequately identified all relevant primary studies?
- **MA_1**: Have you used snowballing?
- **MA_2**: Have you performed pilot searches to train your search string?
- **MA_3**: Have you selected the most-known digital libraries *or* have you made a selection of specific publication venues *or* used broad search engines or indices (*based on the goal of your study*)?
- **MA_4**: Have you compared your list of primary studies to a gold standard or to other secondary studies?
- **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?
- **MA_6**: Have you used a strategy for systematic search string construction?
- **MA_7**: Has an independent expert reviewed the search process?
- **MA_8**: Have you used tools to facilitate the review process?
- **MA_9**: Have you evaluated search results and documented the outcomes?

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?

TV_3: Have you identified primary studies in multiple languages?
- **MA_1**: Is their number expected to be high compared to the population?

TV_4: Were the full texts of all 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?

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?
- **MA_2**: Have you used summaries of candidate primary studies to guarantee the correct identification of all duplicate articles?

TV_6: Have you included/excluded grey literature?
- **MA_1**: Does the decision to include or exclude grey literature comply with the goals of the study *and* the availability of sources?

TV_7: Have you adequately performed study inclusion/exclusion?
- **MA_1**: Have you used systematic voting?
- **MA_2**: Have you performed random screening of articles among authors?
- **MA_3**: Have researchers discussed the inclusion or exclusion of selected articles in case of conflict?
- **MA_4**: Have the inclusion exclusion criteria been documented explicitly in the protocol?
- **MA_5**: Have the authors discussed the inclusion/exclusion criteria *and* revised them after pilots, or by experts' suggestions after review?
- **MA_6**: Have you prescribed a set of decision rules for study inclusion/exclusion?
- **MA_7**: Have you defined quality thresholds for inclusion/exclusion?
- **MA_8**: Have you performed sensitivity analysis?
- **MA_9**: Have you quantified experts' disagreement with the kappa statistic?

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?
- **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?

TV_9: Have you chosen the correct variables to extract?
- **MA_1**: Has the choice of variables been discussed among authors to guarantee that the research questions can be answered?

TV_{10}: Are the studies in your dataset published in a limited set of venues?
- **MA_1**: Have you used snowballing?
- **MA_2**: Have you included grey literature (if this does not affect TV_6)?
- **MA_3**: Have you manually scanned selected venues to check if they publish articles related to your secondary study?

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?

TV_{12}: Does the quality of studies guarantee the validity of extracted data?
- **MA_1**: Have you focused your search process on quality venues only?
- **MA_2**: Have you used article quality assessment as inclusion criterion?
- **MA_3**: Have you assessed the validity of primary studies and their impact using statistics?

TV_{13}: Is there data extraction bias in your study?
- **MA_1**: Have you involved more than one researcher?
- **MA_2**: Have you identified experts' disagreement with kappa statistic?

MA3: Have you performed pilot data extraction to test agreement between researchers? (*Not applicable if MA1 is no*)
MA4: Have you used experts *or* external reviewers' opinion in case of conflicts? (*Not applicable if MA1 is no*)
MA5: Have you performed paper screening to cross-check data extraction?
MA6: Have you used a keywording of abstracts? (*Applicable only in mapping studies*)

TV14: Have you performed statistical analysis?
MA1: Does your data extraction record quantitative data and if yes, does answering your research questions imply the use of statistics?

TV15: Have you selected a robust initial classification schema?
MA1: Have you selected an existing initial classification schema?
MA2: Have you continuously updated the schema, until it becomes stable and classifies all primary studies in one or more classes?

TV16: Is your interpretation of the results subject to bias or is it as objective as possible?
MA1: Have you performed pilot data analysis and interpretation?
MA2: Have you conducted reliability checks (e.g., post-SLR surveys with experts)?
MA3: Have you used a formal data synthesis method?
MA4: Have you performed sensitivity analysis?
MA5: Have you used the scientific quality of primary studies when drawing conclusions?

Research Validity

TV17: Is your process reliable/repeatable?
MA1: Have more than one researcher been involved in the process?
MA2: Have you made all gathered data publicly available?
MA3: Have you documented in detail the review process in a protocol?

TV18: Have you chosen the correct research method?
MA1: 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?
MA2: Have you developed a protocol, monitored the process for deviations, and accurately reported any (if existed)?

TV19: Do the answers to your research questions guarantee the accomplishment of your study goal?
MA1: Have the authors discussed *and* brainstormed on if the research questions holistically cover the goal of the study?
MA2: Is your study and research questions well-motivated?
MA3: Have you consulted target audience for setting up your goals?

TV20: Does your study have substantial related work, so that you can compare and discuss findings?
MA1: Have the authors discussed *and* brainstormed to reach possible interpretations of the findings, due to the absence of related studies?

TV21: Were you familiar with the research field before performing the review?
MA1: 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?

TV22: Are the results of your study generalizable?
MA1: Do your findings comply with those of existing studies?
MA2: Have you used a broad search process w/o an initial starting date?

3. Usage Scenario 1: How Authors can Mitigate Threats

We advise authors to use the checklist and the classification schema provided in this chapter to improve the validity of their study following a number of steps. First, the authors should create a *dedicated section for threats to validity* in both the study protocol and the study report (final manuscript). Second, this section should be *organized according to categories of threats* (e.g., by following the proposed classification schema or another established one). Third, *all threats should be checked whether they pertain to the study.* Finally, for all identified threats, either *appropriate mitigation action should be explicitly reported or an acknowledgement should be made that the threat is not (fully) mitigated*.
To facilitate the aforementioned steps, in the rest of this section (3.1 to 3.3) we present references to representative exemplary mitigation activities from the literature. Finally, in Section 4.4, we summarize the mitigation actions that can be applied in each phase of the secondary study execution.

3.1 Mitigating Threats to Study Selection Validity

Construction of the search string 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 a very limited number (thus missing some relevant studies). A mitigation strategy that covers a wide range of activities is provided by Shanin et al. [50], in which the authors have complemented automated searching in digital libraries with manual search on specific venues that are considered as important to the domain of the secondary study. In addition, the authors have used snowballing (both backward and forward) to decrease the chances of missing articles, i.e., they searched the references of the iden-

tified articles or papers that cite the identified articles for candidate articles they may have missed.

Selection of Digital Libraries 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 missing of relevant studies. As a response to this threat, Garces et al. [24] opted to select the most adequate databases for their search. Based on the criteria discussed by Dieste and Padua [18], they opted for using six databases: namely ACM Digital Library, IEEE Xplore, ScienceDirect, Scopus, Springer, and Web of Science. According to Dyba et al. [21] and Kitchenham and Charters [31], these publication databases are the most relevant sources in the computer science area.

Selection of publication venues 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 that the research aims at high quality studies only. The consequence of this threat is missing relevant studies. A rigorous process for selecting high-quality and relevant publication venues has been discussed in the recent bibliometrics study on top-scholars and institutes [29]. In particular, the authors have selected publication venues based on their relevance to software engineering, their specificity (e.g., architecture, maintenance, etc.), and their average number of citations per month in Google Scholar. Nevertheless, it is also crucial to pilot the searches and compare the obtained studies against a golden standard. An exemplar application of this practice is provided by Jabangwe et al. [27], where the authors have developed the golden standard set by creating an initial validation through Google Scholar, by identifying relevant papers to seminal works (i.e., mostly cited ones) of the secondary study domain.

The ***selection of an arbitrary starting year*** as a starting point for performing the search process can lead to missing studies prior to that date. In order for this decision to not be considered as a threat, it should be clear why such a choice does not influence the results. For instance, according to Li et al. [37], after 2010 there were at least 15 studies published per year focusing on technical debt management, which is a big leap compared with the years before 2010. One reason for this could be that the MTD workshop was initiated in 2010 and this workshop raised the attention on TD and the awareness of managing TD. Therefore, future secondary studies on technical debt could use 2010 as a starting year, without considering this choice as a threat to validity. If such a justification cannot be claimed researchers should consider shifting the starting year earlier.

Problems of the search engines within digital libraries are characterized as ***Search engine inefficiencies*** (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. A tentative mitigation action for this threat is the use of bibliography management tools (e.g. JabRef, Zotero, etc.) for further filtering the large corpus of retrieved articles, based on the desired fields. This mitigation action, although it does not reduce the amount of effort required for data collection, it ensures

the consistency of data collection. A discussion on this is provided by Penzenstadler et al. [44].

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. If the intended scope of the study is indeed narrow there might be no reason to mitigate this threat, as in the case of Santos et al. [49] that focus on action research (i.e., rather young empirical method, that is still under-employed compared to more established ones, e.g., case studies, experiments, etc.) in software engineering. However, alternative strategies could be the inclusion of grey literature, or the execution of broader searches.

Exploring studies written in a specific language (e.g., ***Missing non-English papers***) can lead to the omission of important studies (or number of studies) written in other languages. This threat, exists in almost any secondary study that considers primary studies written in English, since most of them list it as an exclusion criterion. To our opinion this consist a threat only in cases that a very active community publishes high-quality papers in a domain, in languages other than English. A way to evaluate the risk that this threat poses is to assess the number of studies written in non-English languages compared to the population of the research corpus, regardless of the language.

Papers whose full-text is not available cannot be processed (i.e., ***Papers inaccessibility***). If this number is large, the set of retrieved studies might be limited / not representative. As a mitigation action for this threat that is however questionable in terms of generalizability, Magdaleno et al. [39], refer to asking access to the papers through email, directly from the authors. This threat is not very common, since most academic institutes have institutional access to most digital libraries. In case there is no such access, other sources (e.g., research social media, personal websites, etc.) can be used for retrieving a copy, as well as personal contact to the authors by email.

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. For example, Ampatzoglou et al. [6] suggested the merging of multiple versions as one study. In the field of software engineering, a common practice among researchers is to publish their early research results in conference proceedings to get quicker feedback from the research community and as a means for evolving and maturing their work. In many cases a publication to a software engineering journal chronologically follows and includes the results reported in the conference proceedings. In these cases, only the journal article can be added to the set of primary studies without the risk of missing relevant information.

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. For more details on such discussions see the paper of Montalvillo et al. [41]. On the other hand,

if the authors are interested in focusing only on top quality venues (e.g., [9], [23]), then grey literature should be omitted form the searching space.

Study inclusion/exclusion bias 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. As illustrative mitigation action for study inclusion/exclusion, Yang et al. [58] suggest the following strategies: (a) set a group of inclusion and exclusion criteria for study selection, which can be provided as a basis of an objective selection process; (b) considering the possible different interpretation and understanding of selection criteria by the researchers, a pilot selection has to be conducted before the formal selection to guarantee that the researchers reached a clear and consistent understanding of the selection criteria; and (c) two researchers need to conduct the study selection independently in at least in one round of selection, and discuss / resolve any conflicts between their results, to mitigate personal bias in study selection.

3.2 Mitigating Threats to Data Validity

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. The small sample size can be mitigated by broadening the searching space [3], but this decision must comply with the goals of the study and the research area of interest. Additionally, according to Barreiros et al. [11] the small sample size threat is mitigated if the quality of the obtained studies (although low in quantity) is high. Based on the findings of this study, existing secondary studies parse from less than 10 papers to more than 500 primary studies. The mean value is 90 primary studies, whereas 2.5% of our sample includes studies with less than 10 papers and 9.5% of the studies have considered more than 200 papers.

Data from primary studies that are ***highly heterogeneous*** are not easy / safe to synthesize, since such a process is prone to involve a high degree of subjectivity. The mitigation actions that are reported as relevant to this threat are the careful construction of the search string [2], based on the PICO strategy proposed by Kitchenham et al. [33] that takes into account the population, intervention, comparison, and outcomes of the review. Such an approach aims at identifying only the most relevant publications, by limiting the chances for a heterogeneous dataset. Additionally, Nguyen-Duc et al. [43] suggested the development of a data extraction form based on the research questions to ensure that collected data will be as homogenous as possible.

The ***variables*** that have been ***chosen to be extracted*** might threaten the validity of the results, since they might not fit answering the research questions. Additionally, they are prone to researchers' bias. The best practice that can be used for mitigating this threat is the extraction of variables based on the set of research questions and their beforehand mapping. An exemplary way of mapping variables to research

questions is provided by Galster et al. [23], in which the authors list the extracted variables, and inside a parenthesis they denote the research question that can be answered by using them.

Publication bias refers to cases where the majority of primary studies are identified in a specific publication venue. If the majority of primary studies stem from a single workshop, the likelihood of biasing the dataset (the values recorded for every study), and thereof the results, based on the beliefs of a certain community, is rather high. To avoid publication bias, extended and broad searches (e.g., Google Scholar, Scopus, etc.) are encouraged [36], whereas another alternative would be the inclusion of grey literature [54] (e.g., blogs, websites, etc.). Nevertheless, we need to note that both these mitigation actions should be treated with caution, since in specific types of studies, they pose more significant threats to validity. For example, the inclusion of grey literature might hurt the quality of primary studies.

Examining ***data that lack relations*** might hinder reaching a conclusion. A tentative solution to this threat is the application of quality assessment as a criterion for study inclusion or exclusion. In particular, Nguyen-Duc et al. [43] have assessed the quality of the studies in terms of rigor, credibility, and relevance by using the checklist of Dyba and Dingsoyr [21]. An alternative schema for evaluating rigor and relevance for empirical studies has been proposed by Ivarsson [26]. In particular, on one hand, rigor is evaluated based on the description of the context, the empirical design, and the validity discussion. On the other hand, relevance is assessed based on subjects, context, scale, and used research method [31].

Another type of publication bias is the ***Validity of the primary studies***, which suggests 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. The two most common mitigation actions related to this threat are: (a) the use of quality thresholds as an exclusion criterion [1] (e.g., rigor and relevance checklist [21], [26]), and (b) the inclusion of high-quality venues, based on well-defined criteria [29] (see Section 4.1).

Data extraction bias 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. The specific threat to validity is one of the most common ones in software engineering. Therefore, a variety of mitigation actions have been linked to it. The most common ones are: (a) the involvement of more than one researcher in the process and the continuous assessment of their level of agreement (e.g., using Fleis's kappa [43]), (b) the piloting through random sampling [28], and (c) the use of keywording from abstracts [45]. A special type of data extraction bias is the ***Quality assessment subjectivity*** 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. Similarly, ***Data extraction inaccuracies***, refer to cases when 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. Finally, ***unveri-***

fied data extraction refers to the situation in which data are not validated by external reviewers, or have not been subject to internal review. Since all the above threats fall in the generic data extraction bias threats, their mitigation can be achieved by applying the same mitigation actions.

In some designs it is ***not possible to perform statistical analysis***. For example, in cases that all extracted data items are categorical. This threat can be mitigated during the selection of variables to be extracted, when the selection of numerical data can be opted (see above). Nevertheless, as noted by Engström [22], qualitative data analysis methods are equally important to quantitative analysis. Therefore, using solid qualitative analysis methods mitigates the lack of statistical analysis.

Primary studies inconsistent classification is valid for secondary studies that aim at developing a classification schema (usually mapping studies). A similar threat is the ***Construction of attribute framework***. While constructing this framework, the authors define a set of possible values for the attributes (i.e., variables) that are used to characterize each primary study. If the selected values are not discrete and comprehensive then the data extraction can result to an insufficient dataset. In case a classification schema is already in place, ***Robustness of initial classification*** is applicable to secondary studies that rely upon it. A common mitigation while performing the classification of primary studies 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. Actions that can be used for avoiding the aforementioned threat are: (a) the piloting of data extraction to test the classification schema or the attribute framework—Cornelissen et al. [16] evaluated the usefulness of the attribute framework and measured the degree to which the attributes in each facet coincide; (b) the use of an existing and established classification schema—e.g., Hasselberger et al. [25] used the project manager competence development framework; (c) the use of experts' opinion—Kosar et al [35] have relied upon the opinion of a DSL expert for obtaining a coarse-grained classification that could offer a broader picture of the field.

Researcher bias refers to potential bias that 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. To mitigate this threat, vivid discussion among authors of the studies is encouraged, by a variety of studies. Furthermore, Nair et al. [42] advise the execution of reliability checks, the execution of pilot interpretations are proposed by Khurum et al. [30], whereas Penzenstadler et al. compare results with existing studies [44].

3.3 Mitigating Threats to Research Validity

Repeatability 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. The key practice for

boosting the repeatability of a secondary study is the development and the public sharing of a review protocol (e.g., [22]). Other good practices are the involvement of more than one researcher in the process (e.g., Yusifoğlu et al. [59] involved two authors in both data collection and data analysis) and the adoption of well-known guidelines—most studies follow the guidelines of Kitchenham and Charters [31] or of Petersen et al. [45].

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. A discussion on the proper way for selecting the research method for a secondary study is provided by Kitchenham et al. [32]. For example, broad topics should be approached through mapping studies, whereas more specialized ones through SLRs.

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, despite the use of the guidelines of Petersen [45]) threaten validity, since some important aspects might be compromised. In such cases a strong argumentation should be set. For example, Galster et al. [23], deviated from the data extraction guidelines of Kitchenham and Charters [31] and adopted the strategy suggested by Brereton et al. [13].

Coverage of Research Questions refers to the formulation of research questions that do not adequately fulfill the goal of the secondary study. Possible reasons are setting a very generic goal, or the improper decomposition of the goal into questions. The most common best practice for resolving this threat is the use of the GQM approach that has been introduced by Basili et al. [12]. Also, brainstorming among authors [5] and the consultation of experts [4] are highly advisable.

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. Therefore, in our opinion, the only option is the intuitive validation and discussion of the obtained results. A best practice for this is the brainstorming between the authors and possible external experts.

In some cases secondary studies are performed by non-expert researchers that are ***unfamiliar with the research field***. 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, inability to reason about the findings, etc. A tentative best practice for this is the thorough studying of the literature and the detailed comparison of findings. According to Mc Donnell [38] senior researchers should be included in the data analysis and interpretation of the results of secondary studies.

Generalizability threats refer 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***. The mitigation actions that have been linked to generalizability threats are the use of broad searches [19], and the comparison to state-of-the art and related studies [53].

3.4 Mapping Mitigation Actions to Secondary Studies Activities

* In front of each mitigation action the code refers to the id of the threat to validity, followed by the id of the mitigation action in the checklist.

Fig. 4. Mitigation actions that can be applied in each step of the Secondary Study Design Process

To put the application of the aforementioned mitigation actions in context, we assign mitigation actions to activities of secondary studies design processes—see Figure 4. In particular, at the first level (framed font) we present the phases for performing secondary studies as suggested by Kitchenham and Charters [31], and then the corresponding activities (bold font). The used activities are selected as the union of the activities presented in the five studies suggesting guidelines for performing secondary studies [8, 17, 10, 31, and 45]. Being as inclusive as possible in the selection of activities (i.e., by using the union of activities) guarantees that any author will be able to identify the activity that he intends to perform in the figure, regardless of the followed guidelines. In the third level we list the mitigation actions that can be performed in each step. We note that the reporting phase of the secondary studies is omitted since no threats can arise at that stage. However, the step is of paramount importance, in the sense that it includes the reporting of the threats to validity per se.

4. Usage Scenario 2: How Reviewers can Appraise Validity

In this section we illustrate the scenario in which a secondary study needs to be evaluated, either by a reviewer or by a reader of the study, for the purpose of scientific review before publication or for evaluating its validity before usage, respectively. In particular, the evaluation of validity of a secondary study based on the classification schema and the checklist can be performed using two parts of the manuscript: (a) the threats to validity section, and (b) the study design section. We first examine if the threats are classified / organized into sensible categories in the threats to validity section. Subsequently we check if *all threats* to validity are discussed in the threats to validity section, or if *some of them* (or some mitigation actions) are only discussed while reporting the study design.

To illustrate this scenario, we consider a sample of 5 secondary studies that have been performed by the authors of this chapter (and other co-authors). We note that the evaluation provided below does not reflect upon the quality of the published studies, and the trustworthiness of the results, but only focuses on the way that the threats to validity are reported. The five evaluated secondary studies are listed below in chronological order:

[S1] A. Ampatzoglou, and I. Stamelos, “Software engineering research for computer games: A systematic review”, *Information and Software Technology*, Elsevier, 2010 (Ampatzoglou and Stamelos, 2010).

[S2] A. Ampatzoglou, S. Charalampidou, and I. Stamelos, “Research state of the art on GoF design patterns: A mapping study”, *Journal of Systems and Software*, Elsevier, 2013 (Ampatzoglou et al., 2013).

[S3] M. Galster, D. Weyns, D. Tofan, B. Michalik and P. Avgeriou, "Variability in Software Systems—A Systematic Literature Review," *Transactions on Software Engineering*, IEEE Computer Society, 2014 (Galster et al., 2014).

[S4] A. Ampatzoglou, A. Ampatzoglou, A. Chatzigeorgiou, and P. Avgeriou, "The financial aspect of managing technical debt: A systematic literature review", *Information and Software Technology*, Elsevier, 2015 (Ampatzoglou et al., 2015).

[S5] E. M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, M. Galster, and P. Avgeriou, "A mapping study on design-time quality attributes and metrics", *Journal of Systems and Software*, Elsevier, 2017 (Arvanitou et al., 2017).

In Table 1, we present the classification of threats to specific categories in the Threats to Validity section. From Table 1, we can observe that even for studies that come from the same group of authors (or at least overlapping ones), the classification of the threats is not uniform, or it is sometimes completely omitted. Also, we note that for the two studies that are reporting categories, the classes are similar, and quite close to the classification schema reported in Section 2.1. Based on this analysis, reviewers of studies [S1], [S3], and [S5] could point out to authors to either use an established classification schema or come up with their own custom schema. Authors of [S3] should be asked to include an explicit section on validity threats. Reviewers of studies that use custom classifications schemas can encourage the authors to precisely and accurately define them, if they have not done so (which is not the case for [S2] and [S4]). Not all authors need to use an existing schema, but it is crucial that they thoroughly define the types of threats.

Table 1. Classification of Threats into Categories

Study ID	Dedicated Section	Classification of Threats to Validity
[S1]	YES	No categorization
[S2]	YES	Construct Validity. Defined as threats during study design
		Internal Validity. Defined as threats occurring during data collection
		External Validity. Referring to threats when generalizing to population
		Conclusion Validity. Referring to possibly incorrect conclusions (e.g., missing relations, or wrongly extracted relations)
[S3]	NO	No categorization
[S4]	YES	Threats to identification of primary studies
		Threats to data extraction
		Threats to generalization of results
		Threats to conclusions
[S5]	YES	No categorization

Proceeding to a more in depth analysis of reported threats, Table 2 presents which of the threats to validity listed in Section 2.2 have been identified by the five specific studies, how they have been mitigated (the code MA_x of the mitigation action of the corresponding threat TV_y in Table 2), and where (i.e., threats or study design section) they are reported. The rows of the table correspond to a specific threat, the columns to the 5 examined papers, while each cell denotes the corresponding miti-

gation action. A blank cell implies that either the threat is not identified, or it does not apply to the specific secondary study. In case no mitigation action has been taken for a specific threat then we mark it only as identified (ID), but not mitigated. Threats to validity that are discussed in study design (mitigated or not), but not in the "Threats to Validity" section are marked with italics.

Table 2. Identified Threats to Validity

Checklist Question	**[S1]**	**[S2]**	**[S3]**	**[S4]**	**[S5]**
TV_1: Has your search process adequately identified all relevant primary studies?	MA_3 *MA_5*	*MA_3*	*MA_2* MA_3 *MA_5* *MA_6* *MA_9*	*MA_2* *MA_3* *MA_4* *MA_6*	*MA_2* MA_3 *MA_4* *MA_6*
TV_2: Were primary studies relevant to the topic of the review published in several different journals and conferences?	*MA_1*		*MA_1*		
TV_3: Have you identified primary studies in multiple languages?					
TV_4: Were the full texts of all identified primary studies accessible from the researchers					
TV_5: Have you managed duplicate articles?	*MA_1*	*MA_1*	*MA_1*	*MA_1*	*MA_1*
TV_6: Have you included/excluded grey literature?			*MA_1*		*MA_1*
TV_7: Have you adequately performed study inclusion / exclusion?	*MA_3* *MA_4*	*MA_3* *MA_4*	*MA_2* *MA_3* *MA_4* *MA_5*	*MA_3* *MA_4*	*MA_3* *MA_4*
TV_8: Is your sample size large enough so that the obtained results can be considered valid?	*MA_1* *MA_2*	*MA_1*	*MA_1* *MA_2*	*MA_1*	*MA_1*
TV_9: Have you chosen the correct variables to extract?		*MA_1*	*MA_1*		
TV_{10}: Are the primary studies in your dataset published in a limited set of venues?					ID
TV_{11}: Do you expect to identify relationships in your dataset?					
TV_{12}: Does the quality of primary studies guarantee the validity of extracted data?		*MA_1*	MA_1		MA_1
TV_{13}: Is there data extraction bias in your study?		*MA_1* *MA_2*	MA_1 *MA_5*	MA_1	MA_1
TV_{14}: Have you performed statistical analysis?			*MA_1*		*MA_1*
TV_{15}: Have you selected a robust classification schema?	MA_1	MA_1		MA_1	
TV_{16}: Is your interpretation of the results subject to bias or is it as objective as possible?		ID	MA_1	MA_1	MA_1

Checklist Question	[S1]	[S2]	[S3]	[S4]	[S5]
TV_{17}: Is your process reliable/repeatable?	*MA1* *MA2* *MA3*	*MA1* *MA3*	*MA1* *MA3*	*MA1* *MA3*	MA_1 *MA2* MA_3
TV_{18}: Have you chosen the correct research method?		*MA1*	*MA2*		
TV_{19}: Do the answers to your research questions guarantee the accomplishment of your study goal?	*MA2*	*MA2*	*MA2*	*MA2*	*MA2*
TV_{20}: Does your study have substantial related work, so that you can compare and discuss findings?					
TV_{21}: Were you familiar with the research field before performing the review?	MA_1	MA_1	MA_1	MA_1	MA_1
TV_{22}: Are the results of your study generalizable?	*MA2*	ID	*MA2*	ID	

From Table 2 we can observe that the selected studies are covering the majority of the possible threats to validity. Nevertheless, 80.7% of the mitigation actions of studies are only discussed as part of the study design and not the threats to validity section. Although the level of validity for the studies is high, the reporting of the threats is somehow limited. This hinders the evaluation of how threats to validity are considered and mitigated and undermines the overall validity of the studies. In very few cases a threat has been identified without applying any mitigation action, while often more than one action is applied to mitigate a given threat, which implies relatively good management of threats.

Based on this analysis, reviewers could use the proposed classification schema and checklist to encourage the authors: (a) to check whether more threats to validity pertain to their studies, preferably pointing out specific threats that the reviewers have identified; (b) suggest additional mitigation actions for the reported threats that seem more relevant to the study; (c) ensure that all identified threats are mitigated with at least one action; and (d) encourage them to report all the threats identified in the study design, also within the threats to validity section.

5. Recommended Further Reading

We point out three different groups of related work. First, one needs to understand how *threats to validity* are categorized in the *empirical software engineering* field, without focusing on secondary studies. The initial categorization of Cook and Campbell [15] is a fitting starting point, and of course the seminal books by Wohlin et al. [57], Runeson et al. [48], and Shull et al. [52] on experimentation, case study design and empirical SE are also of paramount importance. Second, we advise the interested reader to refer to studies that are related to *the identification and reporting of threats to validity in medical science*, which lies in the heart of the Evidence-Based Software Engineering paradigm. This can provide valuable input for our

field, 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. Indicative readings in this perspective are: [10], [20], [40], [51], and [55]. Finally, to fully comprehend the underlying concepts of this chapter, the readers can refer to the *most common guidelines for performing secondary studies* in the software engineering domain [14], [17], [31], and [45].

7. Conclusions

Threats to the validity of scientific results are inescapable when a particular method or experimental setup is used to collect, analyze and interpret data. In this chapter we have focused on factors that may jeopardize the validity of secondary studies in software engineering. In particular, based on the results of a Systematic Literature Review of secondary studies we have proposed a classification schema, depicting three threat categories (study selection, data and research validity) threats belonging to each category, and the corresponding mitigation actions. To assist authors, reviewers and readers in assessing the rigor of secondary studies we provided a checklist including questions asked to understand if a specific threat is present and corresponding sub-questions to investigate if an appropriate mitigation action has been applied. Finally, we discussed guidelines for identifying and managing threats during the execution of a secondary study and actions for mitigating threats, providing examples and references to the relevant literature.

Secondary studies are a significant driver for the Evidence-Based Software Engineering and often lead to works of major significance that act as reference points in a research topic. Researchers often consult secondary studies to obtain insights to the collective knowledge in a domain and identify opportunities for further research. Ensuring a consistent classification of threats in Systematic Literature Reviews and Mapping Studies and supporting a systematic identification of appropriate mitigation actions can further increase their credibility. Eventually, the proper identification and management of threats can improve the secondary studies' process itself, solidifying the search and selection of primary studies, the extraction of data from the literature and the applied data synthesis.

References

1. A. Ahmad and A. Babar, "Software architectures for robotic systems: A systematic mapping study", *Journal of Systems and Software*, 122, 16-39, 2016.
2. Al-Baik, O., and Miller, J. The kanban approach, between agility and leanness: a systematic review. Empirical Software Engineering, 20(6), 1861-1897, 2015.
3. M.S. Ali, A. Babar, L. Chen and K.J. Stol, "A systematic review of comparative evidence of aspect-oriented programming", *Information and Software Technology*, vol. 52(9), pp. 871-887, 2010.

4. N. S. Alves, T.S. Mendes, M.G de Mendonsa, R.O Spinola, F. Shull and C. Seaman, "Identification and management of technical debt: A systematic mapping study", *Information and Software Technology*, vol. 70, pp. 100-121, 2016.
5. D. Ameller, X. Burgués, O. Collell, D. Costal, X. Franch, and M.P Papazoglou, " Development of service-oriented architectures using model-driven development: A mapping study", *Information and Software Technology*, vol. 62, pp. 42-66, 2015.
6. Ampatzoglou, and I. Stamelos, "Software engineering research for computer games: A systematic review", *Information and Software Technology*, vol. 52(9), pp. 888-901, 2010.
7. Ampatzoglou, A. Ampatzoglou, A. Chatzigeorgiou, and P. Avgeriou, "The financial aspect of managing technical debt: A systematic literature review", *Information and Software Technology*, vol. 64, pp. 52-73, 2015.
8. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies", *Information & Software Technology*, vol.106, pp. 201-230, 2019.
9. E.M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, M. Galster, P. Avgeriou, "A mapping study on design-time quality attributes and metrics", *Journal of Systems and Software*, vol. 127, pp. 52-77, 2017.
10. S.A. Avellar, J. Thomas, R. Kleinman, E. Sama-Miller, S.E, Woodruff, R. Coughlin, T.P.R Westbrook, "External validity: The next step for systematic reviews?", *Evaluation review*, vol. 41(4), pp. 283-325, 2017.
11. E. Barreiros, A. Almeida, J. Saraiva and S. Soares, "A Systematic Mapping Study on Software Engineering Testbeds", *5th International Symposium on Empirical Software Engineering and Measurement*, pp. 107-116, Alberta, Canada, 2011.
12. V. R. Basili and R. W. Selby, "Paradigms for experimentation and empirical studies in software engineering," *Reliab. Eng. Syst. Saf.*, vol. 32, no. 1–2, pp. 171–191, 1991.
13. P. Brereton, B. Kitchenham, D. Budgen, M. Turner, M. Khalilc, "Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain," *Journal of Systems and Software*, vol. 80(4), pp. 571-583, 2007.
14. D. Budgen, P. Brereton, S. Drummond, N. Williams, "Reporting systematic reviews: Some lessons from a tertiary study", *Information and Software Technology*, vol. 95, pp. 62-74, 2018.
15. D. Cook and D. T. Campbell, "Quasi-experimentation: design & analysis issues for field settings". *Boston: Houghton Mifflin*, 1979.
16. Cornelissen, A. Zaidman, A. van Deursen, L. Moonen and R. Koschke, "A Systematic Survey of Program Comprehension through Dynamic Analysis", *IEEE Transactions on Software Engineering*, vol. 35(5), pp. 684-702, 2009.
17. S. Cruzes, T. Dybå, "Research synthesis in software engineering: A tertiary study", Information on Software Technology, vol. 53 (5), pp 440-455, 2011.
18. O. Dieste, D. Padua, "Developing search strategies for detecting relevant experiments for systematic reviews", 1st *International Symposium on Empirical Software Engineering and Measurement* , pp. 215-224, DC, USA, 2007.
19. Ding, P. Liang, A. Tang and H. van Vliet, "Knowledge-based approaches in software documentation: A systematic literature review", *Information and Software Technology*, vol. 56(6), pp. 545-567, 2014.
20. 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, vol. 52(6), pp. 377-384, 1998.
21. T. Dybå, T. Dingsøyr, "Empirical studies of agile software development: a systematic review", *Information and Software Technology*, vol. 50, pp. 833–859, 2008.
22. Engström, M. Skoglund, and P. Runeson, "Empirical Evaluations of Regression Test Selection Techniques: A Systematic Review", *2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement,* pp. 22-31, NY, USA, 2008.

23. M. Galster, D. Weyns, D. Tofan, B. Michalik, and P. Avgeriou, "Variability in Software Systems - A Systematic Literature Review", *IEEE Transactions on Software Engineering*, vol. 40(3), pp. 282–306, 2014.
24. L. Garcés, A. Ampatzoglou, P. Avgeriou, E. Y. Nakagawa, "Quality attributes and quality models for ambient assisted living software systems: A systematic mapping*", Information and Software Technology*, vol. (82), 2017, pp. 121-138, 2017.
25. D. Haselberger, "A literature-based framework of performance-related leadership interactions in ICT project teams", *Information and Software Technology*, vol. 70, pp. 1-17, 2016.
26. M. Ivarsson and T. Gorschek, "A method for evaluating rigor and industrial relevance of technology evaluations", *Empirical Software Engineering*, vol. 16(3), pp.365-395, 2011.
27. R. Jabangwe, J. Borstler, D. Smite and C. Wohlin, "Empirical evidence on the link between object-oriented measures and external quality attributes: a systematic literature review", Empirical *Software Engineering*, vol. 20(3), pp. 640-693, 2015.
28. J. Kabbedijk, C.P. Bezemer, S. Jansen and A. Zaidman, "Defining multi-tenancy: A systematic mapping study on the academic and the industrial perspective", Journal *of Systems and Software*, vol. 100, pp. 139-148, 2015.
29. D. Karanatsiou, Y. Li, E.M. Arvanitou, N. Misirlis, W. E. Wong, "A bibliometric assessment of software engineering scholars and institutions (2010–2017)", *Journal of Systems and Software*, vol. 147, pp. 246- 261, 2019.
30. M. Khurum, and T. Gorschek, "A systematic review of domain analysis solutions for product lines", *Journal of Systems and Software*, vol. 82(12), pp. 1982-2003, 2009.
31. Kitchenham, and S. Charters, "Guidelines for performing systematic literature reviews in software engineering", Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University., (2007)
32. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, S. Linkman, "Systematic literature reviews in software engineering – A tertiary study", Information on Software Technology, Elsevier, vol. 52 (8), pp. 792-805, August, 2010.
33. 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., Elsevier, vol. 51 (1), pp. 7-15, January, 2009.
34. Kitchenham, T. Dybå, M. Jørgensen, "Evidence-based software engineering", Proceedings of the 26th International Conference on Software Engineering (ICSE '04), IEEE, pp 273-281, May, 2004
35. T. Kosar, S. Bohra, and M. Mernik, "Domain-Specific Languages: A Systematic Mapping Study", *Information and Software Technology*, vol. 71, pp. 77-91, 2016.
36. P. Lenberg, R. Feldt and L.G. Wallgren, "Behavioral software engineering: A definition and systematic literature review", *Journal of Systems and Software*, vol. 107, pp. 15-37, 2015.
37. Z. Li, P. Avgeriou and P. Liang, "A systematic mapping study on technical debt and its management", *Journal of Systems and Software*, vol. 101, pp. 193-220, 2015.
38. S. G. MacDonnell, "Invited lighting talks", 23d International Conference on Evaluation and Assessment in Software Engineering, Copenhagen, Denmark, 2019.
39. A.M. Magdaleno, C.M Werner and R.M. Araujo, "Reconciling software development models: A quasi-systematic review", *Journal of Systems and Software*, vol. 85(2), pp. 351-369, 2012.
40. 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, vol 54(1), 2015.
41. L. Montalvillo, and O. Diaz, "Requirement-driven evolution in software product lines: A systematic mapping study", Journal of Systems and Software, vol. 122, pp. 110-143, 2016.
42. S. Nair, J.L. de la Vara, M. Sabetzadeh and L. Briand, "An extended systematic literature review on provision of evidence for safety certification", *Information and Software Technology*, vol. 56(7), 689-717, 2014.

43. Nguyen-Duc, D.S. Cruzes and R. Conradi, "The impact of global dispersion on coordination, team performance and software quality – A systematic literature review", *Information and Software Technology*, vol. 57, pp. 277-294, 2015.
44. Penzenstadler, V. Bauer, C. Caleroand X. Franch, "Sustainability in software engineering: A systematic literature review", 16th International Conference on Evaluation & Assessment in Software Engineering, pp. 32-41, 2012.
45. K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering", In the proceedings of Evaluation and Assessment in Software Engineering, EASE, vol. 8, pp. 68-77. 2008.
46. S. L. Pfleeger and B. A. Kitchenham. 2001. Principles of survey research: part 1: turning lemons into lemonade. SIGSOFT Softw. Eng. Notes 26, 6 November 2001.
47. P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," Empirical Software Engineering, vol. 14 (2), pp. 131–164, December, 2009.
48. P. Runeson, M. Höst, A. Rainer and B. Regnell, "Case study research in software engineering: Guidelines and examples", John Wiley and Sons, 2012.
49. R.E. Santos, F. Q da Silva and C.V de Magalhães, "Benefits and limitations of job rotation in software organizations: a systematic literature review", 20th International Conference on Evaluation and Assessment in Software Engineering, pp. 16.
50. M. Shahin, P. Liang and M.A. Babar, "A systematic review of software architecture visualization techniques", *Journal of Systems and Software*, vol. 94, pp. 161-185, 2014.
51. B. J. Shea, J. M. Grimshaw, G. A. Wells, M. Boers, N. Andersson, C. Hamel, A. C. Porter, P. Tugwell, D. Moher and L. M. Bouter. "Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews", BMC medical research methodology, vol. 7(1), 10, 2007.
52. F. Shull, J. Singer and D.I. Sjøberg, "Guide to advanced empirical software engineering", Springer Science & Business Media, 2007.
53. M. Staples and M. Niazi, "Systematic review of organizational motivations for adopting CMM-based SPI", *Information and Software Technology*, vol. 50(7-8), pp. 605-620, 2008.
54. S. Tiwari and A. Gupta, "A systematic literature review of use case specifications research", *Information and Software Technology*, vol. 67, pp. 128-158, 2015.
55. 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". Journal of clinical epidemiology, 51(12), pp. 1235-1241, 1998.
56. Wohlin, M. Host, P. Runeson, M. Ohlsson, B. Regnell, and A. Wesslen, "Experimentation in software engineering: an introduction", Kluwer Academic Publishers, 2000
57. 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", Journal of Systems and Software, Vol.86 (10), pp. 2594-2610, 2013.
58. C. Yang, P. Liang and P. Avgeriou, "A systematic mapping study on the combination of software architecture and agile development", *Journal of Systems and Software*, vol. 111, pp. 157-184, 2016.
59. V.G Yusifoğlu, Y. Amannejad and A.B Can, "Software test-code engineering: A systematic mapping, *Information and Software Technology*, vol. 58, pp. 123-147, 2015