Journal of Software Engineering Research and Development, 2022, 6:1, doi: 10.5753/jserd.2019.xxx

Assessing the Credibility of Grey Literature: A Study with Brazilian Software Engineering Researchers

Fernando Kamei [**UFPE, IFAL** | ***fernando.kenji@ifal.edu.br***]
Igor Wiese [**UTFPR** | ***igor@utfpr.edu.br***]
Gustavo Pinto [**Zup Innovation & UFPA** | ***gustavo.pinto@zup.com.br***]
Waldemar Ferreira [**UNICAP** | ***waldemar.neto@unicap.br***]
Márcio Ribeiro [**UFAL** | ***marcio@ic.ufal.br***]
Renata Souza [**UFPE** | ***rmcrs@cin.ufpe.br***]
Sérgio Soares [**UFPE** | ***scbs@cin.ufpe.br***]

In recent years, the use and investigations about Grey Literature (GL) increased, in particular, in Software Engineering (SE) research. However, its understanding is still scarce and sometimes controversial, such as interpreting GL types and assessing their credibility. This study aimed to understand the credibility aspects that SE researchers consider in assessing GL and its types. To achieve this goal, we surveyed 53 SE researchers (who answered that they have used GL in our previous investigation), receiving a total of 34 valid responses. Our main findings show that: 1) GL source produced or cited by a renowned source is the main credibility criteria used to assess GL, 2) most of the GL types tend to have a Low to Moderate level of Control and Credibility, 3) there is a positive statistical correlation between the level of Control and Credibility for most GL types, and 4) the different respondent profiles shared similar opinions about the credibility criteria. Our investigation contributes to helping future SE researchers that intend to use GL with more credibility. Additionally, shows the need for future studies to better understand the GL types in SE research.

Keywords: *Grey Literature, Credibility, Empirical Software Engineering, Evidence-Based Software Engineering.*

1 Introduction

Grey Literature (GL) refers to a kind of publication that does not go through a peer-reviewed process before its publication (Petticrew and Roberts, 2006). Some areas of knowledge have used and investigated GL. For instance, in Management, Adams et al. (2016b) investigated how GL could be used with relevance for management and organization studies. In Science of Information (Schöpfel and Prost, 2020), there is an investigation about the term and concept of GL in scientific papers.

In Software Engineering (SE), many researchers interpret GL as any material that was not formally peer-reviewed and published (Garousi et al., 2019). In the last years, SE researchers increased their interest to investigate GL, motivated by the growth of social media and communication channels that SE practitioners use to communicate, exchange problems and ideas (Storey et al., 2017), including, for instance, code hosting websites such as GitHub (Coelho et al., 2020) and communication platforms such as Slack (Stray and Moe, 2020).

In SE, several studies investigated and recognized the importance and usefulness of GL. For instance, Garousi et al. (2016) explored the benefits of GL for Multivocal Literature Reviews, showing what the secondary studies gained when considered GL and what was missed when it was not considered. Other studies (Williams and Rainer, 2017; Rainer and Williams, 2018) investigated the benefits and challenges of using blog content for SE research, and how to improve its use by selecting GL content with more credibility. Despite the increase in investigations in this field, there are some misunderstandings about GL and its diversity types (Tom et al., 2013; Kamei et al., 2021), and how the set of credibility criteria investigated in previous studies (e.g., Williams and Rainer (2017)) could be used and interpreted to the diversity types of GL (Kamei et al., 2021).

According to Adams et al. (2016a), the different types of GL could be classified in terms of the "shades" of grey, which groups GL according to two dimensions: *Control* and *Credibility*. Garousi et al. (2019) explained these dimensions as follows: *Control* is the extent to which content is produced, moderated, or edited in conformance with explicit and transparent knowledge creation criteria. On the other hand, *Credibility* is the extent to which we can determine the producer's authority and knowledge.

In this paper, we begin by studying the different perceptions of SE researchers about GL. We then focused on studying how GL could be assessed considering its different types. For each study, we surveyed Brazilian SE researchers. In the first survey — which was published previously (Kamei et al., 2020) — we investigated how Brazilian SE researchers use GL, focusing on understanding which criteria they employed to assess its credibility as well as the benefits and challenges they perceived. In the second survey (the novel contribution of this paper), we focused on how Brazilian SE researchers that previously used GL perceived the criteria to assess the different GL types according to Control and Credibility.

In the following, we list our main findings (S1 means Survey 1, while S2 means otherwise):

- S1 We identified the main GL sources used by the Brazilian SE researchers;
- S1 We identified several motivations to use (or to avoid) GL;

S1, S2 We identified that the main criteria employed by Brazilian SE researchers to assess GL credibility are: GL source be provided by renowned authors, institutions, companies, or cited by a renowned source;
S2 GL is not widely used as a reference in scientific studies;
S2 We identified different interpretations to assess GL types, showing the importance to consider each type in particular;
S2 We identified for most of the GL types a strong to very strong positive correlations (p-value <= 0.05%) between the perceptions of the level of Control and Credibility;
S2 We did not find a significant correlation (p-value <= 0.05%) between the perceptions of Control and Credibility to GL types when considering the respondent's profile;
S2 We perceived misunderstandings about whether a source type is considered a GL type or not, mainly related to the most classified sources as High Control and High Credibility.

This paper is structured as follows: Section 2 presents the core concepts of this work. Section 3 shows the research questions explored with their rationales. Section 4 exposes the methods employed to conduct, analyze and synthesize the data collected. Section 5 summarizes the answers to the researcher questions (RQ1–RQ4) of the previous investigation (Kamei et al., 2020). Section 6 provides the answers to the research questions (RQ5–RQ6) specifically for this investigation. Section 7 presents the discussions about the findings, lessons learned, and the threats to the validity of this research. Section 8 provides the description and comparison of the related works. Finally, Section 9 exposes the conclusions and future works.

2 Background

Grey Literature (GL) has many definitions. However, the most known is called as Luxembourg definition (Garousi et al., 2019), approved at the Third International Conference on Grey Literature in 1997, that stated: *"[GL] is produced on all levels of government, academics, business, and industry in print and electronic formats, but which is not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body."* Focusing on Software Engineering (SE) research, recently, Garousi et al. (2019) proposed the following definition: *"Grey literature can be defined as any material about SE that is not formally peer-reviewed nor formally published."*

Considering those definitions, they showed a wide concept what would be considered a GL, showing that it can be produced in different ways. However, it may lead to a misunderstanding. For this reason, Adams et al. (2016a) introduced some terms to distinguish the different concepts about grey, including grey literature, grey data, and grey information. The term "grey data" describes user-generated web content (e.g., tweets, blogs, videos). The term "grey information" is informally published or not published (e.g., meeting notes, emails, personal memories). However, SE literature hardly distinguishes these terms. Similarly, we considered all forms of grey data and grey information as GL in our work.

Beyond the GL types, Adams et al. (2016b) classified GL according to "shades of grey". In SE, Garousi et al. (2019) adapted these shades according to three tiers, as shown in Figure 1. In this figure, on the top of the pyramid is the "traditional literature" with scientific articles from conferences and journals. On the rest of the pyramid are what we called as three tiers of GL. These tiers are running according to two dimensions: **Control** and **Credibility**. The first dimension runs between extremes "low" and "higher" and the second runs between extremes "unknown" and "known". The darker the color, the less moderated or edited the source in conformance with explicit and transparent knowledge creation criteria.

Figure 1. The "shades" of grey literature, adapted of Garousi et al. (2019).

Recently, GL was used and investigated in SE research for many purposes. For instance, primary studies explored the GL available on several social media sources used by SE practitioners. For instance, Rainer and Williams (2018) assessed the importance of blog posts to SE research, and Oliveira Oliveira et al. (2021) investigated several Java projects from GitHub to evaluate the developers' skills based on the source code activities.

The presence of GL in secondary studies was notable in the investigations conducted by Zhang et al. (2020) and Kamei et al. (2021) and by the increase in studies based on Grey Literature Reviews (GLR) (e.g., Raulamo-Jurvanen et al. (2017) and Soldani et al. (2018)) and Multivocal Literature Reviews (MLR) (e.g., Garousi et al. (2017) and Saltan (2019)). Explaining these types of study, a GLR is a secondary study that explores the evidence, looking at only GL sources, and a Multivocal Literature is also a secondary study that searches for GL and traditional literature.

Even with this increase in interest in GL, its use is recent in the SE research (Zhang et al., 2020; Kamei et al., 2021). and there are some gaps and different findings of GL in SE research. For instance, Kamei et al. (2021) identified that there is a lack of understanding of what is considered a GL type, and previous studies provide different criteria to assess GL credibility (Kamei et al., 2020; Williams and Rainer, 2019).

3 Research Questions

In this section, we stated our research questions and the rationale for their purposes.

RQ1: *Why do Brazilian SE researchers use grey literature?*

Rationale: Recently, SE practitioners have relied on social media and communication channels to share and acquire knowledge (Storey et al., 2017). On the one hand, some researchers try to take advantage of its use in SE research. For instance, Rainer and Williams (2018) explored the benefits and challenges of blog articles as evidence in SE research. On the other hand, some concerns (e.g., lack of detail and lack of empirical methods) related to GL could make SE researchers skeptical about their credibility (Rainer and Williams, 2019). In this broad question, we intend (i) to understand if Brazilian SE researchers are using GL and, if so, (ii) what motivates them to use, or if not, (iii) the reasons that lead to not using GL.

RQ2: *What types of grey literature are used by Brazilian SE researchers?*

Rationale: According to Adams et al. (2016a), GL has many forms, from traditional mediums such as question & answer websites and blogs to more dynamic mediums such as Telegram and Slack. For this reason, Bonato (2018) emphasized the importance of exploring the GL definition and its types for each research area. There is a lack of understanding of GL types, precisely that the Brazilian SE researchers used. This research question sought to investigate what Brazilian SE researchers often use GL sources. A better understanding of the GL types could guide future research in this area.

RQ3: *What are the criteria Brazilian SE researchers employ to assess grey literature credibility?*

Rationale: Software Engineering research uses GL sources, such as data provided by practitioners retrieved from several social media and communication channels. However, as GL is, by nature, a not peer-reviewed source, SE practitioners are free to share their thoughts using social media, for instance, without worrying about methodological concerns. Thus, it is essential to assess GL sources to ensure the selected GL is appropriate for the study. Answering this question will help us understand the credibility criteria that Brazilian SE researchers consider.

RQ4: *What benefits and challenges Brazilian SE researchers perceive when using grey literature?*

Rationale: According to Storey et al. (2014), the SE research community has increased its interest in GL since the widespread presence of SE professionals using social media and communication channels. For instance, exploring the Stack Overflow, Zahedi et al. (2020) found some trends and challenges in continuous SE that researchers could better explore. In this question, we are interested in understanding the (i) benefits and (ii) challenges that researchers may face when resorting to GL. Answering this question is essential to understand the potential benefits and challenges of using GL more broadly by researchers.

RQ5: *How do SE researchers prioritize a set of criteria to assess grey literature credibility?*

Rationale: In our first investigation (Kamei et al., 2020), we provided a set of criteria used by Brazilian SE researchers to assess GL credibility. Previous literature (Williams and Rainer, 2019) also identified another set of criteria. In this question, we focused on understanding the importance of those criteria to assess GL credibility.

RQ6: *What is the perception of Brazilian SE researchers about the different types of Grey Literature according to the perspective of Control and Credibility?*

Rationale: Due to the diverse nature of the GL types, some studies suggested that GL needs to be assessed in different ways (Garousi et al., 2019). For this reason, Adams Adams et al. (2016b) classified its types according to the shades of grey. This classification is based on two dimensions: *Control* and *Credibility*. Control refers to the rigor with which a source is produced. Credibility, the extent to which the knowledge and producer authority can be determined. Nevertheless, this understanding and classification are still confused. This research question sought to understand how Brazilian SE researchers commonly perceived the GL types according to the (i) Control and (ii) Credibility.

4 Research Methods

In this work, we followed (Linåker et al., 2015), aiming to use a survey methodology for data collection. This data was collected from a group of people sampled from a large population. We conducted two surveys. The first (Survey 1) aimed to understand the Brazilian SE researcher's perceptions about GL. The second (Survey 2) investigated only the Brazilian researchers from the first survey who answered that they used GL.

In the following sections, we detailed the procedures used to conduct Survey 1 with participants of a flagship conference of SE in Brazil (Section 4.1). Then, we present the procedures used for Survey 2 that focused on the researchers that have experience using GL (Section 4.2). Finally, we provide the methods used for the analysis of both surveys (Section 4.3).

4.1 Survey 1: Initial investigation with the Brazilian SE researchers

In Survey 1, we intended to gather a broad perception of GL used by Brazilian SE researchers, focusing on understanding the motivations to use (or avoid), the types of GL used, the benefits and challenges, and the criteria used to assess its credibility.

4.1.1 Survey Design

We conducted our survey with participants of the 10th Brazilian Conference on Software: Practice and Theory (CBSoft), the largest Brazilian software conference with many SE researchers' participation. It includes well-established and specialized satellite SE conferences in its domain. Our population comprehends SE researchers potentially interested in using GL in their research. We chose our sample using non-probabilistic sampling by convenience (Baltes and Ralph, 2021).

Before sending the final survey version, an experienced researcher (Ph.D. SE researcher with more than 15 years of experience in research) reviewed our draft. We also conducted a pilot study by randomly selecting two participants and explicitly asking for their feedback. We received feedback suggesting changing the order and re-writing some questions to make them more understandable to the target population.

We obtained the contact of all the 252 participants, asking the conference's general chair whether s/he could share this information with us, which s/he gently provided.[1]

We used two approaches to invite the researchers to answer our questionnaire. First, we placed posters on the event's walls and tables with a brief description of the work and the link to the online survey. Second, we sent the actual survey to the 250 remaining participants of the event. In the invitation email, we briefly introduced ourselves, presented the research's purposes, highlighted that the invite was to the participant of the CBSoft, and the link to the online survey. We also mentioned that the participant was free to withdraw at any moment, and all information stored was confidential.

The survey was open for responses from September 26th to October 11th, 2019. We received a total of 76 valid answers (30.4% response rate). We did not consider the pilot survey answers.

4.1.2 Survey Respondents

Among the survey respondents, 48.7% have a Ph.D., 31.6% have a Master's, 2.6% are graduate specialization, 14.5% have a Bachelor's degree, and 2.6% are undergraduates. Among them, 72.4% are men, and 27.6% are women. Table 1 presents the demographics' information about the respondents and their experience using GL or not. This table shows that most respondents with Ph.D. and Master's degrees answered that they were using GL.

[1] In the period of this research, the Brazilian General Data Protection Law was not yet officially published.

Table 1. Demographics information of the Survey 1 respondents.

Gender	Level of course	Used GL	Not used GL
Woman	Doctorate	5	5
Man	Doctorate	24	3
Woman	Master	4	2
Man	Master	15	3
Woman	Expert	1	1
Man	Expert	0	0
Woman	University graduate	0	2
Man	University graduate	2	7
Woman	Technical education	0	0
Man	Technical education	0	0
Woman	High school	1	0
Man	High school	1	0

4.1.3 Survey Questions

Our survey had 11 questions (three were required, nine of which were open). We used different questions flow for those who used GL (did not answer question 10) from those who did not (answered only questions 1 to 4 and questions 10 and 11). Table 2 presented the questions covered in this survey.

4.2 Survey 2: Investigating Brazilian SE researchers that use Grey Literature

In this survey, we intended to do a follow-up survey to collect perceptions only with the Brazilian SE researchers from Survey 1, who answered that they have previously used GL. We focused on the perceptions of the different GL types concerning the dimensions of Control and Credibility.

4.2.1 Survey design

Using a non-probability sample by convenience (Baltes and Ralph, 2021), we invited by email once again the 53 researchers that participated in our Survey 1 and mentioned the use of GL.

We first drew our questionnaire and improved it through the conduction of three sequential steps: 1) A pilot study with five Ph.D. SE researchers; 2) Another SE researcher specialist assessed the questionnaire; and 3) Received feedback of a participant relating a problem in the first hours after opening the survey. For this reason, we closed the survey to stop receiving answers.

Then, we deleted all answers previously received and sent a new questionnaire version to the researchers. We opened the survey for answers from February 10th to March 4th, 2021. We received a total of 34 valid answers (64.1% response rate). We did not consider the pilot survey answers.

4.2.2 Survey Respondents

In this survey, as we retrieved our sample from the previous one who answered that they had used GL, we did not ask the same questions (e.g., gender, academic degree). Instead, we collected information about their experience in SE research and using GL in scientific articles.

Table 2. Questions covered in the Survey 1.

#	Question	Type of question	Options of answers (for closed questions)	Required?	RQ
Q1	What is your e-mail?	Open	-	No	-
Q2	What is your gender?	Open	-	Yes	-
Q3	Please list the highest academic degree you have received.	Closed	High school, Technical education, University graduate, Expert, Master's degree, Doctorate.	Yes	-
Q4	Have you used grey literature? If you never used, go to question *Q10*.	Closed	Yes, No.	Yes	RQ1
Q5	What sources of grey literature did you use?	Open	-	No	RQ2
Q6	In which conditions *do you use* grey literature?	Open	-	No	RQ1
Q7	In which conditions do you do *not use* grey literature?	Open	-	No	RQ1
Q8	Could you list any *benefits* in using grey literature?	Open	-	No	RQ4
Q9	Could you list any *challenges* in using grey literature?	Open	-	No	RQ4
Q10	If you answered 'no' in question four, please state why did you *never use* or *avoid* use grey literature?	Open	-	No	RQ1
Q11	What would be a *reliable source* of grey literature for you?	Open	-	No	RQ3

The respondents' profile of our survey was composed of 76.5% of professors or researchers and 23.5% of undergraduates. Regarding SE research experience, 55.9% of the respondents had more than ten years. Considering the experience using GL, 47% had conducted between 2 and 5 scientific studies using GL, although 26.5% were unable to answer.

4.2.3 Survey Questions

Our second survey had ten questions (six were required, and four were open). Table 3 presents the questions covered in this survey. Before question 4, we produced and included a video[2] to summarize and explain the "shades of GL" according to the level of Control and Credibility.

4.3 Data Analysis and Synthesis

In both surveys, we employed a mixed-method approach based on both qualitative (Section 4.3.1) and quantitative (Section 4.3.2) methods to analyze data. We used a *qualitative* approach when we were interested in questions about "what" and "how" and a *quantitative* analysis using descriptive statistics to discuss frequency and distribution and correlation analysis between the dimensions of Control and Credibility to each GL type. We describe these methods in the following.

4.3.1 Qualitative analysis

We used a qualitative approach based on the thematic analysis technique (Braun and Clarke, 2006). This process involved three SE researchers with previous qualitative research experience (one Ph.D. student (R1) and two Ph.D. professors (R2–R3)) for both surveys.

We performed an agreement analysis with the codes and categories generated by each researcher using the Kappa statistic (Viera and Garrett, 2005) to Survey 1. The Kappa value was 0.749, indicating a Substantial Agreement level, according to the Kappa reference table (Viera and Garrett, 2005). For Survey 2, we do not calculate Kappa due to the analysis process that occurred with the researchers working together.

Figure 2 presents a general overview of the process employed. In the following, we detailed the procedure used to analyze all the answers (adapted from Pinto et al. (2019)) of both surveys, showing the differences employed in each survey research:

1. *Familiarizing with data:* The process starts with two independent researchers reading the answers of the survey respondents, as expressed in Figure 2-(a).
2. *Initial coding:* Then, for Survey 1, two independent researchers (R1 and R2) individually analyzed and added codes. For Survey 2, the researchers analyzed, discussed, and coded together (R1 and R2, into a dotted box). We used a post-formed code, so we labeled portions of text that expressed the meaning of the excerpts without any previous pre-formed code. The initial codes are temporaries, since they still need refinement. We refined the emerged codes throughout all the analyses. An example of coding is present in Figure 2-(b).
3. *From codes to categories:* Here, we already had an initial list of codes. For Survey 1, two researchers individually conducted this process (R1 and R2). For Survey

[2]Video explaining the "shades of GL" (in Portuguese): `https://youtu.be/hGMkVXIApRO`

Table 3. Questions covered in the Survey 2.

#	Question	Type of question	Options of answers (for closed questions)	Required?	RQ
Q1	What is your occupation?	Closed	Professor/Researcher, Student (M.Sc. or Ph.D.), Other (open).	Yes	-
Q2	How many years of experience did you have conducting SE research?	Closed	Until 1 year, From 1 and 3years, From 4 to 6 years, From 7 to 9 years, 10 years or more.	Yes	-
Q3	How many scientific studies have you conducted using GL as source of evidence?	Closed	I do not know, No one, Only one, From 2 and 5, From 6 and 10, More than 10.	Yes	-
Q4	We are aware that the *level of Control* varies from source to source. For this reason, we ask you to consider your experience more frequent in relation to each source type in relation to the *Control* dimension of the production.	Closed	Source types: {adapted from Maro et al. (2018); Level of Control: I did not consider it as a GL type, Low Control, Moderate Control, High Control, No opinion.	Yes	RQ6
Q5	Please, explain what did you consider to classify each source type with the *Control criteria* presented in Question 5.	Open	-	No	RQ6
Q6	We are aware that the *level of Credibility* varies from source to source. For this reason, we ask you to consider your experience more frequent in relation to each source type in relation to the *Credibility* dimension of the production.	Closed	Source types: {adapted from Maro et al. (2018); Level of Credibility: I did not consider it as a GL type, Low Credibility, Moderate Credibility, High Credibility, No opinion.	Yes	RQ6
Q7	Please, explain what did you consider to classify each source type with the *Credibility criteria* presented in Question 7.	Open	-	Yes	RQ6
Q8	Considering a GL source with important information to your research, would you include a GL source if it is produced by/with.	Closed	Choices for Credibility criteria: Be produced by a renowned author, Be produced by a renowned institution, Be produced by a renowned company, Be cited by others renowned sources, Describe the methods of collection, Cites an academic reference, Cites a practitioner source, Presents information with rigor, Presents empirical data; Choices for answers: No opinion, No, Yes.	Yes	RQ5
Q9	Could you cite any additional potential aspect to assess the credibility of a GL source that was not mentioned before?	Open	-	No	RQ6
Q10	We are planning to conduct a future research about Quality Assessment in Grey Literature. Please, could you inform your mail to future contact?	Open	-	No	-

2, this process occurred with two researchers working together (R1 and R2). This process begins to look for similar codes in the data. We grouped the codes with similar characteristics in broader categories. Eventually, we also had to refine the categories identified, comparing and re-analyzing them in parallel, using an approach similar to axial coding (Spencer, 2009). Figure 2-(c) presents an example of this process.

4. *Categories refinement:* Here, we have a potential set of categories. For both surveys, in a consensus meeting between R1 and R2 (Figure 2-(d)), the categories were evaluated and solved the disagreements of interpretation for evidence that supported or refuted the categories found. We also rename or regroup some categories to describe the excerpts better there. In cases where disagreements remained, we invited a third researcher (a Ph.D. professor) to review and solve them for both surveys.

4.3.2 Quantitative analysis

We based our quantitative investigation on three samples: (i) We used the answers from 76 SE researchers to answer RQ1; (ii) We used the answers from 53 researchers that mentioned using GL to answer RQ2, RQ3, and RQ4; and (iii) We used the answers from 34 to answer RQ5 and RQ6.

For the descriptive statistics, we highlighted that one answer of a respondent could be related to more than one category found. In the investigations related between the GL types and the dimensions of Control and Credibility, we present it into boxplots to show the differences of interpretations of each GL type.

We used Spearman's rank correlation coefficient for the correlation analysis of the Control and Credibility perceptions for each GL type. Then, we transformed the answers related to the level of Control and Credibility (Low, Moderate, High) into non-linear scales: Low = 0, Moderate = 50, and High = 100.

For the quantitative data analysis, we used R language and Python. This last, with the support of Google Colab[3].

5 Previous Results

In this section, we summarized the findings of our first study to present answers to RQ1–RQ4. To understand these research questions, consider reading the previous study (Kamei et al., 2020).

To each RQ, we summarized the categories in tables with the total number of occurrences of a given category in the column "#". Two critical observations are required: 1) The researchers may have reported more than one answer per question, which may happen to be grouped into different categories; and 2) Some questions are not required. Thus, the overall results might not reach 100% of respondents.

[3]https://colab.research.google.com

RQ1: Why do Brazilian SE researchers use grey literature?

In our Survey 1, we identified 53 SE researchers using GL for research purposes. Focusing on understanding better why and how SE researchers are using GL or avoiding its use, we asked questions that included the motivations to use GL or reasons to avoid it. In the following, we present a summary of the (i) *motivations to use GL* and (ii) and the *reasons to avoid or never use GL*.

(i) Motivations to use

Table 4 presents the identified SE researchers' motivations to use GL. In this table, the first column describes the motivation identified, followed by the number of respondents related to the category and the percentage associated with the total of SE researchers that used GL (n=53). In the following, we briefly describe some motivations.

Table 4. Motivations to use GL.

Motivation	#	%
To understand the problems	28	52.8%
To complement research findings	12	22.6%
To answer practical and technical questions	10	18.9%
To prepare classes	4	7.5%
To conduct government studies	1	1.9%

To understand problems was the most cited motivation to use GL, where several researchers noted the use of GL for some reasons: to understand or investigate a new topic, or to search for something to solve problems, or to acquire specific information to deepen the knowledge.

To complement research findings was the second most cited motivation, mentioned when the knowledge gained from the traditional literature is not enough for the investigation. For instance, a researcher noted the use of GL to complement the findings of a Mapping Study.

To answer practical and technical questions was the third most cited motivation, related to the necessity to understand the state of the practice in SE.

Other motivations were mentioned but to less extent, such as ***To prepare class*** and ***To conduct government studies***.

(iii) Reasons to avoid/never use

Even though several motivations to use GL were identified, 50.9% of SE researchers (27/53) ***avoid using*** GL as a reference or to reinforce some claims in scientific studies. We also found some researchers that ***never used GL*** (23/76 occurrences, 30.3%) to any research situations. We used this value to analyze the extent of each category about reasons to never use GL. Of the 23 respondents that never used GL, only 15 answered the reason. Table 5 presents the summary of the findings for this question. In the following, we briefly describe the reasons to avoid GL.

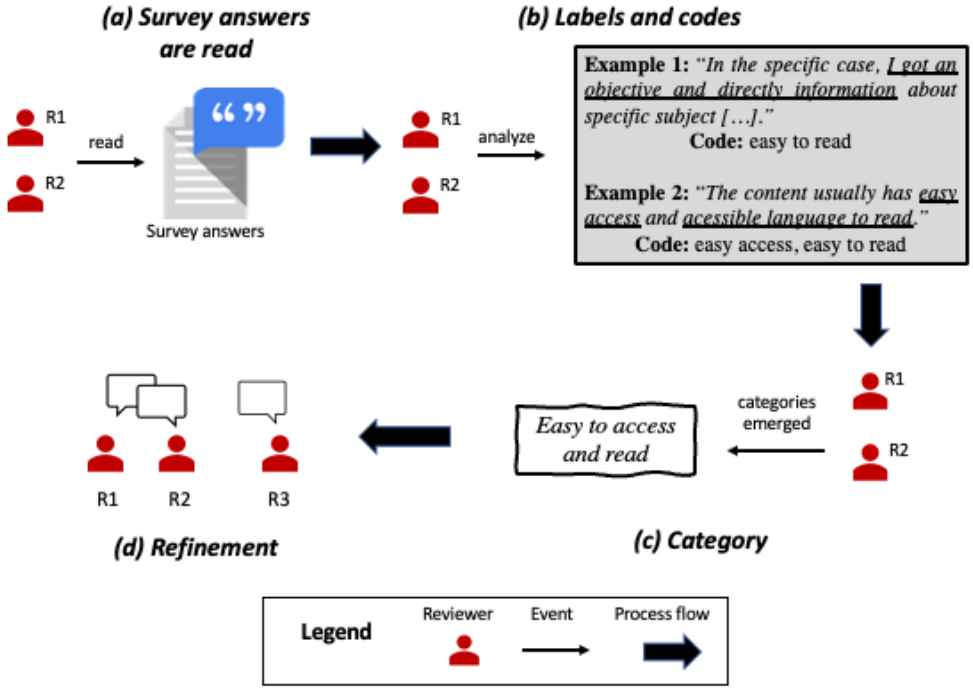

Figure 2. Example of a coding process used to analyze the questionnaire answers.

Table 5. Reasons to avoid/never use GL.

Reason	#	%
Lack of reliability	6	26%
Lack of scientific value	3	13%
Lack of opportunity to use	3	13%

Lack of reliability was the main reason that SE researchers mentioned not to use GL. This is related to the lack of rigor in which GL sources are written and published, which affects its credibility.

Lack of scientific value was another category mentioned, where the researchers were afraid that the use of GL would weaken a research paper when submitted to the peer-review process.

Lack of opportunity to use was related to the nature of research previously conducted and because GL is recent in the context of SE.

Summary of RQ1: Brazilian SE researchers use GL motivated mainly to understand new topics, find information about practical and technical questions, and complement research findings. However, some researchers avoid GL, particularly as references in scientific papers, due to its lack of reliability and scientific value.

RQ2: What types of grey literature are used by Brazilian SE researchers?

In this question, we explored the GL sources used by the 53 SE researchers that mentioned use GL. Table 6 listed these sources. In the following, we briefly present some of our findings.

Q&A websites was the most common source mentioned, used to interact with other users, create content, post comments, and assess the content. Some examples of sources mentioning Q&A websites were Stack Overflow and Quora.

Blog post was the second most common category found. Blogs from renowned practitioners and from companies that produce a diversity of material and content of SE and software development, in general, were mentioned.

Technical reports were mentioned for SE researchers that used technical experience, reports, and surveys derived from industry and national and international research groups.

Companies websites provided by Google, Facebook, and ThoughtWorks, containing information regarding their technologies, methods, and practices, were mentioned as sources used. Some researchers said browsing these websites to find news to help decision-making about a specific technology.

Table 6. GL sources used by SE researchers.

Source	#	%
Q&A websites	16	30.2%
Blog posts	15	28.3%
Technical reports	14	26.4%
Companies websites	8	15%
Preprints	5	9.4%
Books/Book chapters	5	9.4%
Software repositories	4	7.5%
Videos	3	5.7%
Magazine articles	3	5.7%
News articles	2	3.8%

Summary of RQ2: Brazilian SE researchers are using several GL sources. The most common are *Q&A websites*, *blog*

posts, *technical reports*, and *companies websites*.

RQ3: What are the criteria Brazilian SE researchers employ to assess grey literature credibility?

In this research question, we explored the answers in one open-ended question the criteria of how the SE researchers assess GL credibility. Table 7 summarized our findings. In the following, we briefly describe the criteria identified.

Renowned authors was the criterion most cited, in which SE researchers considered the author's experience and reputation concerning the topic. For instance, Martin Fowler was cited as a notorious software engineer with much knowledge.

Renowned institutions was another crucial criterion, where SE researchers assess if renowned institutions or renowned research groups provided the GL content.

Cited by others was a criterion mentioned to express those researchers that considered as a trusted source cited by others (studies or people).

Renowned companies was a criterion identified that consider relevant when renowned software industries or portals produce the GL source.

Table 7. Criteria to assess GL credibility.

Criteria	#	%
Renowned authors	15	28.3%
Renowned institutions	14	26.4%
Cited by a renowned source	8	15%
Renowned companies	7	13.2%

Summary of RQ3: Whoever produces GL's content, whether made by a person, institution, or company since the producer is considered renowned, is a significant credibility criterion.

RQ4: What benefits and challenges Brazilian SE researchers perceive when using grey literature?

In this research question, we explored the benefits and challenges on the GL use mentioned by SE researchers. Table 8 summarizes the benefits and Table 9 the challenges. In the following, we briefly describe some of them.

Table 8. Benefits of the use of GL.

Benefit	#	%
Easy to access and read	16	30.2%
Provide a Practical Evidence	13	24.5%
Knowledge acquisition	13	24.5%
Updated information	6	11.3%
Advance the state of the art/practice	5	9.4%
Different results from scientific studies	3	5.7%

Table 9. Challenges of the use of GL.

Challenge	#	%
Lack of reliability	34	64.2%
Lack of scientific value	15	28.3%
Difficult to search/find information	6	11.3%
Non-structured information	6	11.3%

(i) Benefits

Easy to access and read was the most common benefit mentioned, mainly because most GL sources are open access, are quickly recovered by free search engines, and the contents are usually easy to read.

Empirical evidence was another essential benefit mentioned, showing that GL provides evidence from the SE industry to understand the state of the practice.

Knowledge acquisition was mentioned as a benefit, as GL allows expanding knowledge with different information from what is usually obtained in traditional literature.

Updated information was mentioned because the production of GL content happens fast compared with traditional literature, mainly related to technical content.

Advance the state of the art/practice was mentioned due to the importance of GL to understand better the industry and to provide evidence to find relevant gaps in the practice.

Advance the state of the art/practice was mentioned due to GL's importance in understanding the industry better and providing evidence to find essential gaps in the practice.

Different results from scientific studies was mentioned because some researchers considered GL essential to provide additional knowledge not yet available in the research area.

(ii) Challenges

Lack of reliability was the main challenge the researchers perceived, where some questioned the reliability of the data retrieved from GL.

Lack of scientific value was the second category most cited. Some researchers mentioned that they did not feel comfortable using GL as a reference in scientific works due to the research community's lack of recognition of this source.

Difficult to search/find information in GL sources was perceived as a challenge due to the diversity of sources. Each source has its structure and manner to provide access to the content, and it is not easy to replicate the study that used GL.

Non-structured information was mentioned due to the lack of a writing pattern and a large variety of formats in which the GL sources are published, making it difficult to find information, for instance, using an automatic process.

Summary of RQ4: We found several benefits, the most common was that the GL's content is easy to access and read, which is important to knowledge acquisition, mainly about providing practical evidence derived from SE practitioners. The most cited challenges were using GL in scientific research due to the lack of reliability and scientific value.

Table 10. Prioritized criteria to assess GL credibility.

Criteria	#	%
Renowned authors	30	88.2%
Renowned institutions	30	88.2%
Cited by a renowned source	27	79.4%
Cites academic source[a]	26	76.5%
Present empirical data[a]	26	76.5%
Renowned companies	25	73.5%
Cites practitioner source[a]	16	47.1%
Rigor in presenting information[a]	12	35.3%
Describe the methods of collection[a]	6	17.6%

[a]Proposed in Williams and Rainer (2019)

6 Results

In this section, we present answers to RQ5 and RQ6, both research questions answered by the investigation of Survey 2.

RQ5: How do SE researchers prioritize a set of criteria to assess grey literature credibility?

In our second survey, we asked 53 researchers to prioritize the importance of a set of criteria to assess GL credibility. These criteria were derived from our first investigation and found in Williams and Rainer (2019) study. We received answers from 34 SE researchers. Table 10 presents the result of the ranking prioritization of credibility criteria, revealing that essential criteria perceived by SE researchers are: GL source be provided by a *Renowned authors*, *Renowned institutions*, or *Cited by a renowned source*.

We also investigated whether the SE researchers have any additional criterion to assess GL credibility not mentioned in the previous survey questions. By analyzing the answers, we did not find any new criterion that was not related to the criteria as earlier presented in Table 10. For instance, some researchers mentioned that the detailed description of the publication context is an important criterion. For this case, we considered that it is already contemplated in *Rigor in presenting information* criterion, previously mentioned by Williams and Rainer (2019). The author's experience with the topic was another criterion mentioned. We considered this criterion related to the *Renowned author's* criterion identified in our first survey.

Summary of RQ5: We assessed the prioritization of credibility criteria identified in our first investigation, in addition to those identified in previous studies. We found that the most used criteria by SE researchers are when the GL is produced by a renowned source, cited by a renowned authority, cites an academic source, and presents empirical data.

RQ6: What is the perception of Brazilian SE researchers about the different types of Grey Literature according to the perspective of Control and Credibility?

Our last research question explored how the researchers perceived the different types of GL concern to the dimensions of Control and Credibility. These dimensions are used to classify the tiers of the "shades of GL." Each dimension could be evaluated into three levels (Low, Moderate, High). Figure 3 presents the results of classifications according to the level of Control, and Figure 4 the results of the level of Credibility.

Even we are investigating different dimensions, interestingly, in some cases, the Figures 3 and 4 presented similar behaviors. For instance, for some GL types (e.g., *blog posts*, *forums/list of discussions*), the Low level was predominantly in both dimensions. We also found similarities concerning the other levels for both dimensions. For instance, some types (e.g., *materials training*, *news articles*, *software repositories*, and *tutorials*) run between Low (1st Quartile) to Moderate (2nd Quartile). Although, for a diversity of cases, the median behavior varied.

We also found differences. For instance, considering the level of Control to *cases/services descriptions* and *guidelines*, the classifications run between Low (1st Quartile) to Moderate (2nd Quartile). In contrast, for the level of Credibility to these GL types, we found outliers on the Low level (1st Quartile) and outliers on the High level (3rd Quartile).

Other classifications caught our attention. For instance, regarding the Control dimension, the opinions about the *magazine articles* are not equalized, as we identified some outliers in both extremes (Low and High). A similar classification we identified related to *guidelines* for the Credibility dimension.

In addition to classifying the levels (Low, Moderate, and High) of the dimensions (Control and Credibility), we offered the possibility to the researcher to choose the options of *"I did not consider it a GL type"* or *"I have no opinion."* We included these options because even previous studies (e.g., Maro et al. (2018)) presented the GL types for SE research; in our previous investigation (Kamei et al., 2021), we identified different interpretations, for instance, in which some types were not considered as GL. Table 11 shows the results of these classifications.

Comparing the findings presented in Table 11 with the information presented in Figures 3 and 4, we perceived that most of GL types classified with High Credibility and High Control were also, many times, considered as not a GL type (e.g., *thesis*, *books/book chapters*, and *patents*). Moreover, we identified that *patents* are still unknown to several researchers.

Rationale to employ classification of each dimension (Control and Credibility)

We asked why the researchers employed the classifications of each GL type according to the Control and Credibility. We identified four main reasons that are summarized in Table 12 and described in the following.

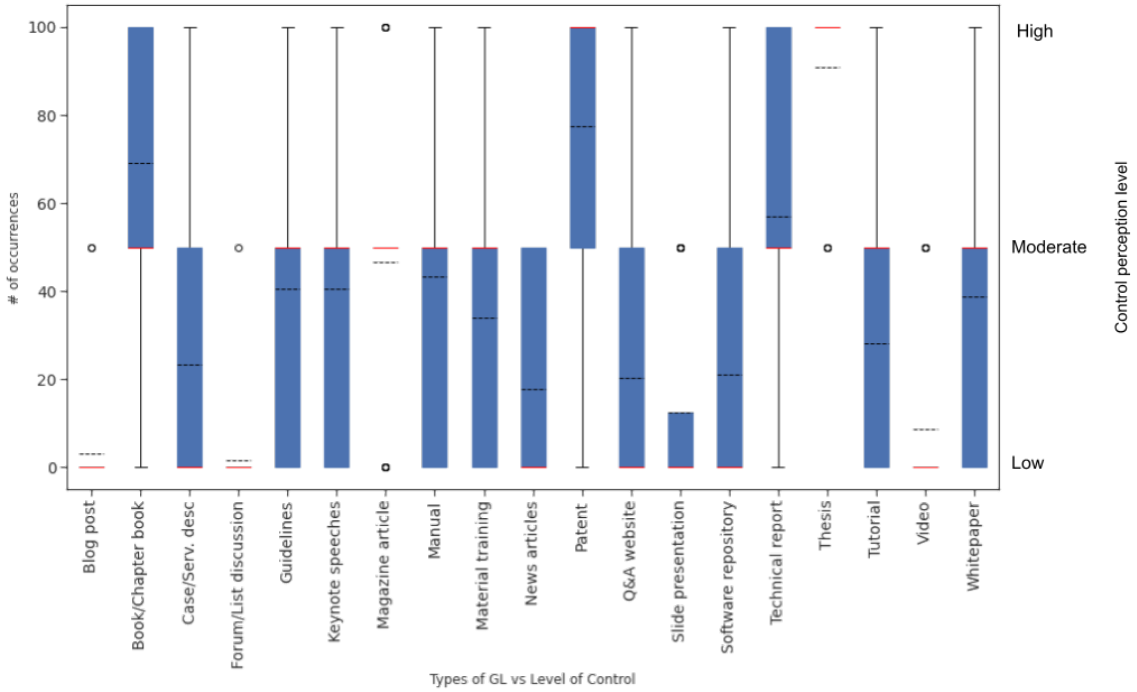

Figure 3. Classification of each GL source type according to the level of Control. Each level of Control indicates: Low = 0; Moderate = 50; High = 100.

Table 12. Reasons to classify GL types according to the level of Control and Credibility.

Reasons	#	%
Rigor	23	67.6%
Producer reputation	14	41.2%
Research experience	13	38.2%
Peer interaction	5	14.7%

***Rigor* (23/34 occurrences)**. Researchers considered the rigor (control) of each source's production, for instance, the degree of formality present. In this regard, one researcher pointed out: *"Technical reports, for instance, present systematic studies with high control (of production)."* This category was also related to the credibility dimension, as one researcher affirmed: *"I consider that credibility is directly related to the rigor of the publication/availability of an artifact."*

***Producer reputation* (14/34 occurrences)**. The producer's reputation was considered an essential criterion to assess Control and Credibility, as one researcher pointed out: *"The credibility relates to who is the author of the material and to the platform being conveyed*. Another one mentioned: *"Depending on the publisher, I can consider high (e.g., Elsevier) or low (e.g., autonomously published book) control. The same applies to news: the credibility of the source influences the level of control regarding stricter editorial control in favor of the integrity of the information."*

***Researcher experience* (13/34 occurrences)**. The own researchers' experience was used to employ the classification. In this regard, one researcher pointed out: *"I thought of the examples for each type that I have used and classified them according to my experience in dealing with each material."* Another one mentioned that: *"I considered what I have read about grey literature."*

***Peer interaction* (5/34 occurrences)**. Another criterion considered for assessing GL Control and Credibility was the users' interactions in GL sources. In this regard, one researcher mentioned: *"Another point is that if I have a lot of people interacting and building the content (such as Q&A websites), I consider that it has a certain control in the final knowledge presented there."* Another one pointed out: *"In general, I consider the control to be higher when there is a peer review in some way, as in the case of theses and Stack Overflow."*

Correlation analysis between the level of the dimensions (Control and Credibility) and each GL type

We conducted our analysis using correlation statistics between the two variables (Control and Credibility) to each GL type using the Spearman coefficient. We interpreted the Spearman coefficient according to Dancey and Reidy (2004). To conduct this analysis, aiming to pair the samples, we removed the answers in which one respondent answered that *"I did not consider it a GL type"* or *"I have no opinion"* to at least one dimension to the same GL type.

Based on the results of Spearman's rank correlation presented in Table 13, we identified 13 GL types (13/19; 68.4%), with correlations that varies from **strong to very strong positive correlations (p-value <= 0.05% of significance).** It indicates that when the Control's level increases, the Credibility tends to increase.

Considering only the group of GL types that presented **less than 95% of significance**, we identified six types. Among these types, 4 out of 6 (*forums/list of discussions*, *cases/services descriptions*, *keynote speeches*, *materials training*) had **moderate correlations**. For the remaining two (*books/book chapters* and *magazine articles*), we identified a **negligible correlations**.

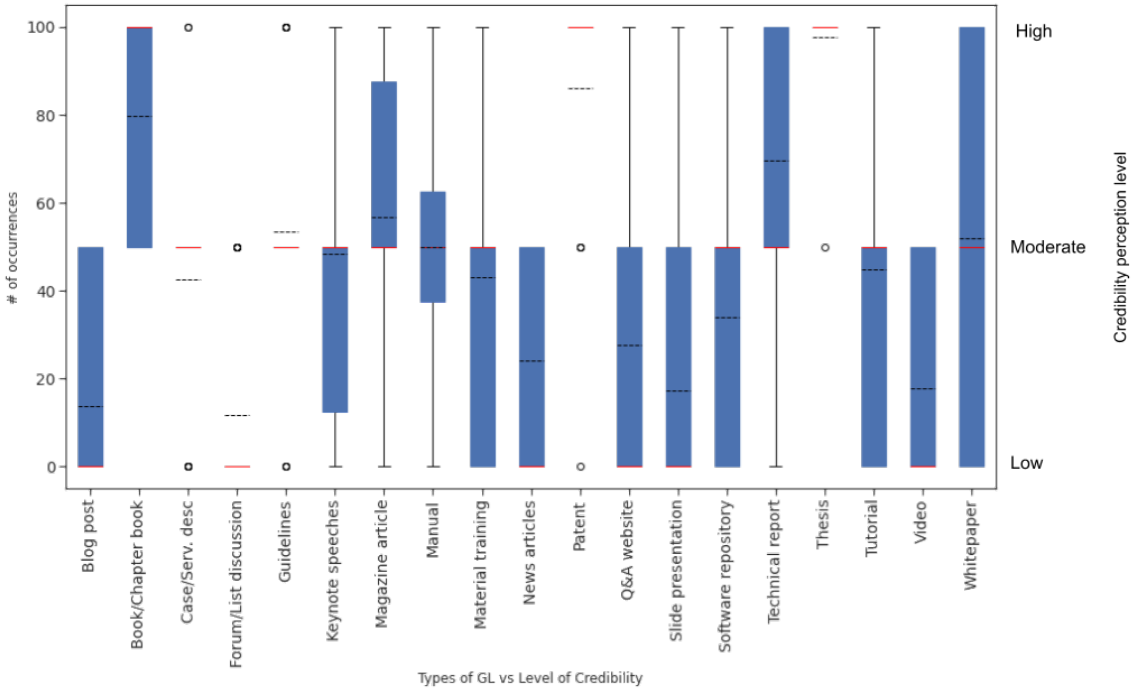

Figure 4. Classification of each GL source type according to the level of Credibility. Each level of Credibility indicates: Low = 0; Moderate = 50; High = 100.

Table 13. Types of Grey Literature: Control and Credibility correlation test. Notes: *Correlation is significant (strong) at the rho >= 0.4 and p-value <= 0.05 level; ***p-value* is not zero (we used three decimal places).

Type of Grey Literature	Spearman coefficient	*P*-value
Blog post	**.441***	.017
Book/Book chapter	.106	.607
Case/Soft. description	.341	.082
Forum/Discussion list	.337	.069
Guideline	**.518***	.004
Keynote speeches	.305	.101
Magazine article	.167	.377
Manual	**.620***	.000**
Material training	.308	.104
News articles	**.525***	.003
Patent	**.550***	.027
Q&A websites	**.656***	.000**
Slide presentation	**.593***	.001
Soft. Repository	**.652***	.000**
Technical report	**.527***	.005
Thesis	**.546***	.013
Tutorial	**.688***	.000**
Video	**.671***	.000**
White paper	**.769***	.000**

Correlation analysis between the level of the dimensions (Control and Credibility) and the respondent profiles

After analyzing our data, a chi-square test of independence was conducted between the respondent profiles and their inclination to answer *"I did not consider it a GL type"* or *"I have no opinion"*. Therefore, we evaluated if the fact that the respondent is a professor or not has any influence in not considering as GL or not having an opinion. Table 14 presents our result.

Table 14. Chi-square test between respondent profiles and (i) Not considered as GL, (ii) No opinion - Control, and (iii) No opinion - Credibility.

Type of GL	i	ii	iii
Blog post	.769	.526	.959
Book/Book chapter	.925	.959	.526
Case/Soft. description	.959	.959	.439
Forum/Discussion list	.959	.999	.959
Guideline	.526	.526	.579
Keynote speeches	.959	.999	.959
Magazine article	.959	.526	.959
Manual	.769	.526	.769
Material training	.959	.999	.769
News articles	.959	.999	.769
Patent	.883	.393	.726
Q&A websites	.769	.526	.959
Slide presentation	.959	.999	.925
Soft. Repository	.769	.999	.769
Technical report	.959	.769	.769
Thesis	.526	.999	.194
Tutorial	.526	.999	.579
Video	.959	.769	.579
White paper	.959	959	.711

As we can see in Table 14, we did not have found a statistically significant association ($p < 0.05$) between respondent profile and their inclination to have no opinion regarding the level of Control and Credibility, or did not consider as a GL type. Therefore, based on our results, we did not reject any null hypothesis, i.e., the respondent profile did not influence their answers, or our sample is not large enough to show this influence.

Table 11. The types of GL in which SE researchers have no opinion regarding the level of *Control* and *Credibility*, or do not consider as GL (✖ GL).

Type of source	Control No opinion	Credibility No opinion	✖ GL
Thesis	0	1	12
Patents	7	10	7
Books/Book chapters	2	1	6
Magazine articles	1	2	3
Case/Serv. desc	1	5	3
Manuals	1	3	3
Materials training	0	3	3
Software repositories	0	3	3
Blog posts	1	3	2
Forums / Lists	0	2	2
News articles	0	3	2
Slide presentations	0	6	2
Keynote speeches	0	2	2
Videos	3	4	2
Technical reports	3	2	2
Q&A websites	1	3	1
Guidelines	1	4	1
Tutorials	0	4	1
White papers	2	5	1

We performed another Chi-square statistical test to discover if the respondent profiles affect results to their opinion on Low, Moderate, or High level of Control and Credibility. For each factor (Control or Credibility) and GL (*blog posts*, *books/book chapters*, etc.), we populated a 2X3 contingency table composed of rows (i.e., respondent profile) and columns (i.e., their opinion as Low, Moderate, or High) variables. Table 15 presents the p-value from the chi-square statistical test for each contingency table.

Table 15. Chi-square test between respondent profiles and (i) Credibility level and (ii) Control level.

Type of GL	Credibility	Control
Blog post	.785	.100
Book/Book chapter	.958	.722
Case/Soft. description	.632	.293
Forum/Discussion list	.720	.557
Guideline	.769	.853
Keynote speeches	.185	.853
Magazine article	.539	.692
Manual	.496	.069
Material training	.316	.690
News articles	**.049**	.205
Patent	.651	.905
Q&A websites	.567	.289
Slide presentation	.478	.157
Soft. Repository	.387	.261
Technical report	.848	.743
Thesis	.746	.844
Tutorial	.132	.707
Video	.755	.894
White paper	.925	.752

Table 15 shows the distribution of the p-values per comparison from each Chi-squared test of independence. As we can see, there is no evidence that different respondent profiles have different opinions. The only exception regards *news articles* credibility. The contingency table (see Table 16) summarizes the results from comparing answers from professors/students and *news articles* credibility. We conclude that professors/researchers think that *news articles* are more believable by analyzing this result.

Table 16. Contingency table from respondent profiles and the levels of Credibility for *news articles*

Respondent profile	Low	Moderate	High
Professors/researchers	7	1	0
Students	8	13	0

Summary of RQ6: We identified similar behaviors when considering the same GL type concerning the two dimensions: Control and Credibility. Most GL types ran between the Low and Moderate levels in these dimensions. We also identified some differences, such as the median of answers for Control were at the Low level and a Moderate level for the Credibility dimension. The production rigor, the producer's reputation, researcher experience, and the permission of peer interaction are the criteria employed by the researchers to assess GL source. Moreover, we found some misunderstandings to consider or not some data sources as GL, mainly related to *thesis*, *patents*, *magazine articles*, and *books/book chapters*. Considering the correlation analysis, we identified that it varied from strong to very strong between Control and Credibility dimensions for most GL types. Our investigation also shows a correlation analysis between the level of Control and Credibility for most GL types, showing that when one dimension increases, the other one tends to increase too. The same happens when the level decrease. Considering the researcher profile, we did not find evidence that different researcher's profiles have different opinions, except for the *news articles*.

7 Discussion

In this section, we discussed each research question, relating them to previous studies (Section 7.1). Then, we discussed some findings out of the scope of the RQs that caught our attention (Section 7.2). We also presented some advice to SE researchers based on the lessons learned with this research and previous knowledge (Section 7.3). Finally, we discussed some threats to the validity of this work (Section 7.4).

7.1 Revisiting findings

In this section, we discussed our findings to each RQ. Even we have addressed the RQ1–RQ4 in our previous study (Kamei et al., 2020), in this work, we included additional discussions and considered other related works not mentioned before.

(RQ1) Motivations to use or reasons to avoid GL

(i) Even our first investigation showed several motivations and benefits in using GL. Our second investigation shows that most researchers avoid its use as a reference in scientific papers.

(ii) We organized the motivations to use GL into five categories. Three of them were similar to previous works. For instance, Rainer and Williams (2019) and Zhang et al. (2020) also discussed the motivation *to complement research findings*. Another related motivation was *to understand problems*, identified in three studies (Rainer and Williams, 2019; Neto et al., 2019; Zhang et al., 2020).

(RQ2) Types of Grey Literature used

We did not find previous primary studies focusing on this research question. We found tertiary studies that investigated the most GL types found in selected studies. For instance, Zhang et al. (2020) identified that the most common GL types used in the list of selected secondary studies were (in order) *technical reports*, *blog posts*, *books/book chapters*, and *thesis*.

Considering the types of GL used by Brazilian SE researchers, the most common are the *Q&A websites* (e.g., Stack Overflow), *blog posts* (e.g., SE firms, such as Netflix, Uber, Facebook), and *technical reports* (e.g., from SEI). Our investigation shows that most of these types are related to SE practice, mainly retrieved from renowned firms or research institutions.

(RQ3) Criteria used to assess Grey Literature credibility

We found several criteria to assess the GL credibility, showing that most of them are related to the GL producer *be renowned (authors, institutions, and companies)*. These criteria caught our attention because we did not find any criterion mentioning to assess the GL content. However, the challenge of *Lack of reliability* identified is related to this, and previous work (Williams and Rainer, 2019) have investigated a set of criteria to assess GL content (e.g., *rigor in presenting information, presenting empirical data, describe the methods of data collection*).

(RQ4) Benefits and Challenges using Grey Literature

We identified some contradictory findings between the benefits and challenges of GL use. They are part of the trade-off between traditional literature and GL nature. For instance, on the one hand, SE researchers mentioned that it is *Easy to access and read* the GL content. On the other hand, they said the *Difficult to search/find information*. Regarding the benefit, it is related to accessing the GL content without paywall restriction and to the informal language usually written. However, these benefits hinder the use of automatic data extraction.

We identified another trade-off, for instance, even the perceived benefit of *Advance the state of the art/practice*, several researchers are avoiding the use of GL due to the challenges of *Lack of reliability* and *Lack of scientific value*. In part, those trade-offs are expected, showing the necessity for further investigations on how to improve the use of GL in SE research. For instance, as we have done in this research.

Even we confirmed some findings of the literature, the main benefit identified (*Easy to access and read*) was not mentioned by previous studies (Williams and Rainer, 2017; Rainer and Williams, 2018, 2019; Garousi et al., 2016). Similarly, it occurred with the challenges. For instance, the *Lack of scientific value* was not identified in previous studies. Even, it was the second challenge most mentioned in our investigation. We informed that the benefits identified in this study are related to our results of a tertiary study (Kamei et al., 2021). Regarding the challenges, some findings in previous works (Zhang et al., 2020; Kamei et al., 2021). For instance, the *Uncertain availability of GL* was not identified in our investigation.

(RQ5) Prioritizing the Criteria to Assess Grey Literature Credibility

This investigation confirmed some findings of Survey 1 (Kamei et al., 2020), showing that the most important credibility criteria are related to the GL source be produced by a *renowned source*. However, using the prioritization criteria, some of these findings contrasted partly because, in Survey 1 results, no criteria were related to assessing the GL content. At the same time, in Survey 2, several SE researchers considered important criteria of *Cites academic source* and *Present empirical data*.

The criteria of citing academic sources, describing the collection methods, and presenting empirical data caught our attention due to the emphasis on applying scientific perspectives to assess GL sources. In our opinion, these criteria are difficult to be used, as we discuss in the following: 1) According to Williams (2018), online articles and blogs produced by SE practitioners rarely mentioned academic sources; 2) GL sources are produced mainly by practitioners (Kamei et al., 2021), and consultant/companies have different manners of expressing than academics one; and 3) Most of the GL sources do not present empirical data. Instead, they are primarily based on their opinions and belief (Rainer, 2017).

(RQ6) Types of Grey Literature vs Dimensions of Control and Credibility

Some findings caught our attention because some GL types run between two and sometimes into three levels of the classification of the dimensions, showing that different interpretations may occur for the same type. Although, the correlation analysis showed a strong correlation between these interpretations for most of the GL types investigated. Considering the respondent's profiles, different from what we expected, our statistical analysis based on the Chi-square test showed that different respondent profiles shared similar opinions about each source type being considered a GL or not and concerning the level of control and credibility.

The criteria used by SE researchers to classify these dimensions are mostly related to the rigor of source, researcher experience, and the interaction permitted for the user to deal with each GL type. Although some of them considered it is challenging to classify considering only the source type,

without a real example to be deeply assessed, as one researcher pointed out: *"(...) the credibility will depend on who produced that content."* Moreover, we perceived that sources (e.g., *technical reports*, *books/book chapters*, *thesis*) produced by companies and institutions mainly were considered with Moderate to a High level of Control and Credibility. In contrast, the sources commonly produced by SE practitioners (e.g., *forums/list of discussions*, *blog posts*, *videos*) have a Low level of Control and Credibility. These findings caught our attention because, in RQ2 results, the most used GL sources runs between Low to Moderate level. It appears that the benefits and the motivations to use GL outweigh the Low level of Control and Credibility presented in these sources.

With these findings, we reinforce the claim of Garousi et al. (2019) that it is complicated to assess the dimensions of Control and Credibility alone. Although they could bring us one direction, other essential criteria include identifying GL's producer and content. For this reason, we advocate that SE researchers use the concept of the "shades of GL" to classify and assess a GL source because it recognizes the different perspectives of the nature of GL, although future investigations to set a limit between tiers of the shades are essential. Beyond that, we claimed the importance of employing objective criteria to assess GL sources and better permit the GL classification according to the shades. Although, as our findings showed, it could be essential to propose intermediate shades between each tier.

7.2 Other discussions

In this section, we discussed some findings and important discussions unrelated to a specific research question. First, we discussed the relations among the researcher's perceptions' of GL. Second, describe the relationship between the credibility criteria and the dimensions of credibility investigated. Lastly, discuss our findings of the perceptions of the different GL types.

Perceptions of Grey Literature

We identified relations between the perceptions of GL, as shown in Figure 5. For instance, we identified some *motivations to use* GL related to some *benefits* identified (slashed line) and some *reasons to avoid* GL with some *challenges* by GL use (dotted line). In what follows, we discussed some of them.

Regarding the *motivation to use* "To complement research findings" is related to the *benefit* of use GL to provide "Different results from scientific studies" as some respondents informed that the inclusion of GL could provide evidence not explored or identified in the research area. Another one is "To answer practical and technical question" related to the benefit of "Practical evidence", which was not perceived using only traditional literature.

The *reasons to avoid* GL and the *challenges* identified are almost the same. Except for the "Lack of reliability" that hinders the replicability of the search for GL. It could be motivated due to the "Non-structured information" of a GL source.

Credibility criteria vs Dimensions of Control and Credibility

The most important criteria identified to assess GL credibility are related to the "Producer reputation" and the "Rigor" presented in the GL source. The first is related to the source be produced by a renowned author, institution, or cited by a renowned source. The second with how the information is presented, for instance, if it describes the methods used to collect the data. Figure 6 presented these criteria.

We also identified some relations between the credibility criteria with some reasons to classify the Control and Credibility dimensions, as shown in Figure 6. The Control (slashed line) is related to the "peer interaction", "producer reputation", and the "rigor". The Credibility (dotted line), their relations are the same as the Control dimension, including the "researcher experience". This last is related to their own researcher experience using GL to assess its credibility.

GL types interpretation

In our second investigation, we found some misunderstanding in interpreting GL types (see Table 11), even though those types were recognized as GL in some previous SE works (e.g., Maro et al. (2018), Zhang et al. (2020)). In the following, we present the most common types that were not considered GL: *thesis* (11/34 occurrences), *patents* (6/34 occurrences), *books/book chapters* (6/34 occurrences), and *magazine articles* (3/34 occurrences). In this regard, for instance, one researcher pointed out: *"I understand that thesis and dissertations are not Grey because external researchers formally assess them."*

We also found in previous studies some contradictions in interpreting a source type as a GL type or not. For instance, while Hosseinzadeh et al. (2018) considered *books/book chapters* as a GL type, the study of Berg et al. (2018) did not. We identified another conflict, for instance, while Neto et al. (2019) considered *thesis* a peer-reviewed source, Rodríguez-Pérez et al. (2018) classified them as GL types. These misunderstandings were also identified in the previous investigation with secondary studies (Kamei et al., 2021).

In our opinion, these misunderstandings reflect on each source's classification regarding Control and Credibility. For instance, for most researchers, *books/book chapters*, *technical reports*, *thesis*, and *patents* were not considered a GL type and related them to a High level of Control and Credibility (Figures 3 and 4). It shows that the peer-reviewed process and grey literature boundary are unclear when considering only the source type.

7.3 Lessons learned

With this investigation and the previous one (Kamei et al., 2020), we showed how GL could contribute to SE research. However, some advice is important to this use could be improved.

For **SE researchers,** our findings highlight to pay attention when searching, selecting, and using grey literature in SE research: 1) Explore the GL sources before using on their research, as there are several types of GL source, to understand

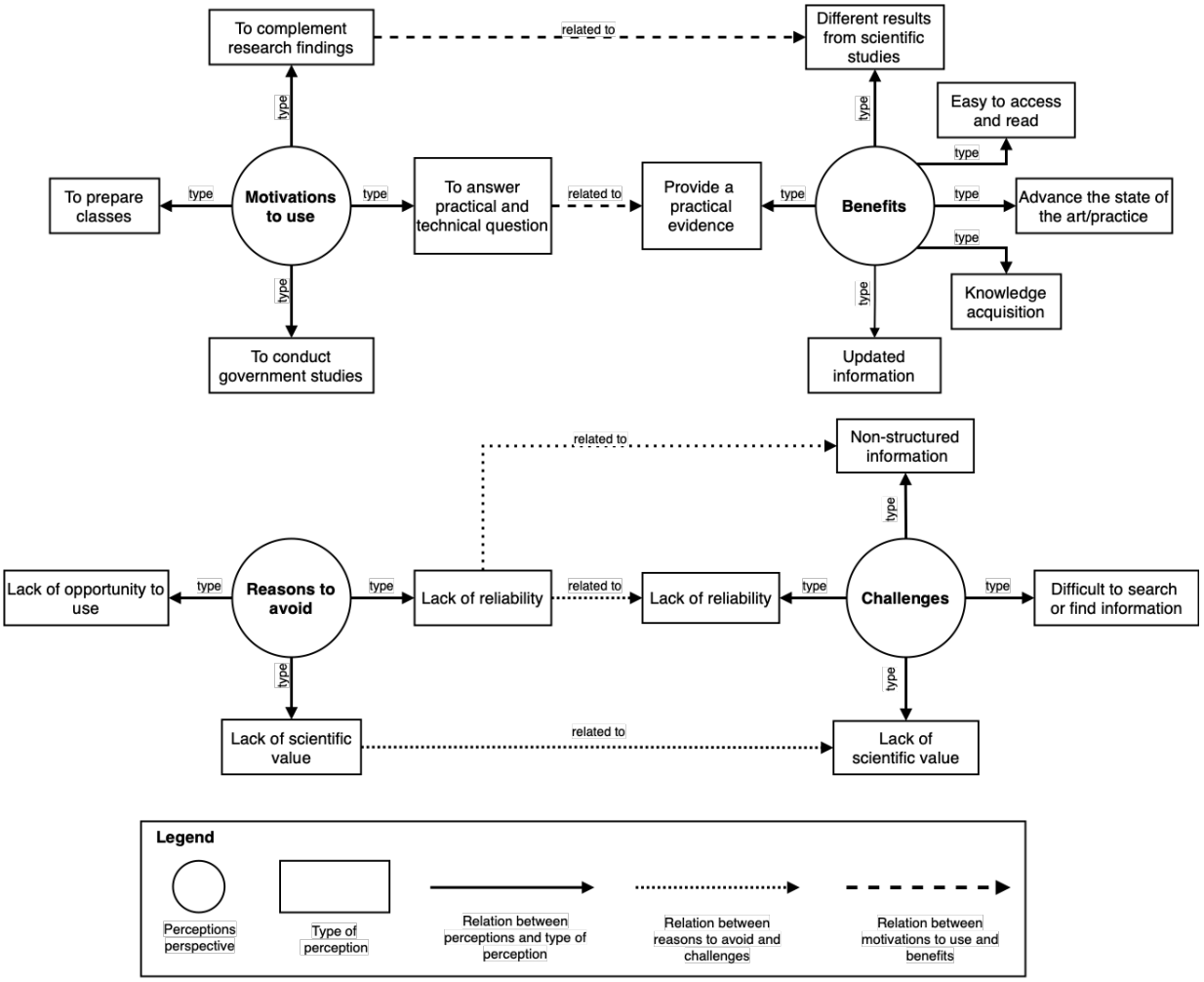

Figure 5. Relationships identified between the Motivations to Use GL with Benefits and the Reasons to avoid with the Challenges.

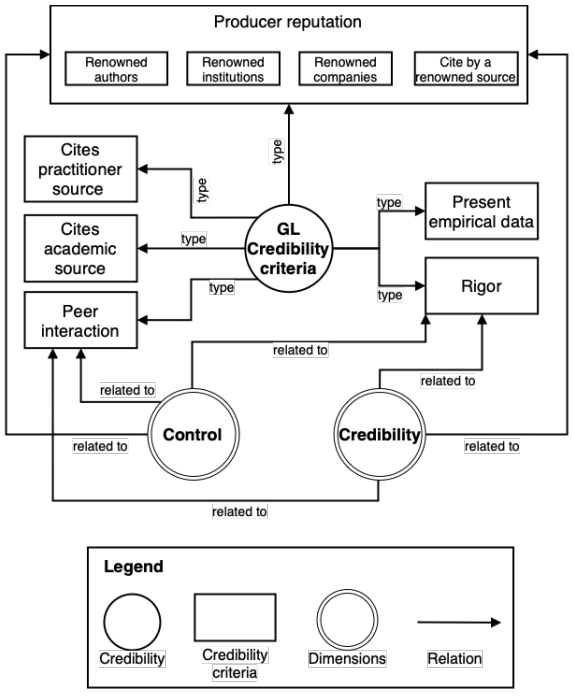

Figure 6. Relationships identified between the Grey Literature Credibility criteria with the Dimensions of Control and Credibility.

what evidence each GL source could provide and could benefit the research and how to retrieve information from them, due to the issues about the difficulty to search for; 2) It is important to the researchers be aware of a set of credibility crite-

ria that could be used to assess GL sources. For instance, by selecting data produced by renowned sources (e.g., authors, institutions) and understanding how each credibility criteria could better fit each type of GL; 3) Another criterion to improve GL credibility could be used, considering the various interpretations for GL assessment related to the Control and Credibility aspects; and 4) Understand how to improve the search for GL using a systematic approach with methods and techniques to better deal with the content, aiming to reduce their lack of reliability.

7.4 Threats to Validity

This section discussed some limitations and threats to validity and what we have done to mitigate them.

Construct validity: Even our efforts to improve our questionnaire, we identified two potential threats in our research: 1) Specifically on the questions that we asked for the participant to classify each source type concerning the Control and Credibility dimensions. We mitigate this, informing the researchers that we know that Control and Credibility vary from source to source, and asked them to consider the most frequent experience for each data source. However, three researchers reported that assessing these GL types' dimensions was difficult without considering the content and the producer. This difficulty may have introduced some bias, and 2) We used a non-probability sample by convenience (Baltes and Ralph, 2021) because we intend to investigate only SE researchers with previous experience in GL use. Then, we surveyed only 53 Brazilian researchers we knew had this experience.

Internal validity: As our investigation used personal interpretation, we may have introduced biases during the data extraction and analysis. We tried to minimize those by using a paired approach with a constant discussion between the researchers and invoking a third researcher to revise the derived codes and categories.

External validity: Our first investigation used a sample of the SE researchers from the largest SE conference in Brazil. In the second investigation, our sample was representative of SE research because we had a 30.4% response rate with a diversity of researchers (1/3 are women, 50% have a Ph.D. in SE, and 30% a Master's). In our second investigation, we conducted our survey with the researchers from the first survey that mentioned they had used GL in SE research. We received 64.1% of response rate. From these, almost 60% are professors or researchers with more than ten years of SE research experience, and most have used GL from 2 and 5 scientific studies. Nevertheless, as we focused on the Brazilian SE research community for both surveys, the findings may not apply to other populations. Although, we used the peer review process during all this research, aiming to improve the external validity to draw general conclusions.

Conclusion validity: Even with 30.4% and 64.1% of response rates in both surveys, we may have lost some important information. For the first investigation, we mitigated this threat by comparing our results with previous studies conducted with different populations, showing that our results showed similarly. Even though we have reached a considerable response rate for the second investigation, our sample was small and focused only on the Brazilian SE researchers' perspective to permit the results' generalization. Another threat is related to the correlation analysis between the dimensions of Control and Credibility to each GL type because we did not explicitly ask this correlation to the respondents.

8 Related works

This section group the related works in studies that explored GL's credibility and quality assessment in SE research. For each study presented, we show the differences concerning our work.

The Grey Literature Review (GLR) conducted by Raulamo-Jurvanen et al. (2017) focused on understanding how SE practitioners choose a test automation tool by investigating opinions and experiences of SE practitioners produced in GL sources. They analyzed the GL source's credibility during the quality assessment according to the number of readers, number of shares, number of comments, number of Google Hits for the titles, and adopting backlinks analysis (a reference comparable to a citation). Our work differs because we provide different findings on assessing GL credibility. Moreover, we also intend to understand the prioritization of a set of criteria identified in previous investigations (Kamei et al., 2020; Williams and Rainer, 2019).

Soldani et al. (2018) conducted another study based on GLR. This study investigated the pains and gains of the use of microservices. They perceived that the traditional literature on the topic is still in the early stage even though companies are working day-by-day with microservices, as witnessed by the considerable amount of GL on the subject. The authors considered a set of criteria of control factors to select GL sources: Practical Experience of the authors (+5 years), Industrial case-study, Heterogeneity (present the information about at least 5 top industrial domains), and Implementation quantity (present detailed information). Our work differs from this because we focused on investigating and providing a set of general criteria that could be used to assess different types of GL sources.

Williams and Rainer conducted two studies to investigate how to improve the quality and credibility assessment of blog articles in SE research. The first study (Williams and Rainer, 2017) examined some criteria to evaluate blog articles to be used as a source of SE research evidence through two pilot studies (a systematic mapping study and preliminary analyses of blog posts). The findings showed some criteria to select a blog article's content (e.g., authentic, informative). The second study (Williams and Rainer, 2019) focused on finding credibility criteria to assess blog posts by selecting 88 candidate credibility criteria from a previous Mapping Study (Williams and Rainer, 2017). Then, to gather opinions on a blog post to evaluate those credibility criteria, they surveyed 43 SE researchers. Some criteria were found, for instance, the presence of reasoning, reporting empirical data, and reporting data collection methods. As discussed in the previous related works, our criteria were not focused on a spe-

cific type of GL. Moreover, our identified criteria are different from Williams and Rainer's, and we tried to understand what each SE researcher considered assessing the different types of GL.

Most recently, we conducted a tertiary study with secondary studies of SE (Kamei et al., 2021) presenting a critical review of GL use in secondary studies. In total, were investigated 446 studies, identifying 126 studies that searched or included GL as a primary source. This finding showed that GL was not widely used in the analyzed studies, although it increased in GL use over the years. The tertiary research explored the benefits, challenges, and motivations to use or avoid GL use. Our work differs from this previous one because we asked the SE researchers directly, different from investigations with published studies, where these questions were not directly explored, leaving the authors the option to include or not that information.

Even though the similarity of these works with our work, there are differences in at least four points: i) We found a different set of credibility criteria: the source needed to be provided by renowned institutions, renowned companies, cited by others, and derived from academia, ii) We did not focus on a specific type of GL source, iii) We explored the experience of SE researchers to understand the perspectives on the credibility of different GL types and how SE researchers assess them, and iv) We investigated a set of prioritization criteria used to assess GL credibility.

9 Conclusions and Future Works

Although the use and investigation of Grey Literature in SE research increased over the last years, they are still recent.

In this work, we reported two investigations based on the Brazilian SE researchers' perspective to present an overview of GL sources usage, potential benefits and challenges of its use, a set of criteria to assess GL credibility, and the perceptions about GL types concerning Control and Credibility criteria. Our main findings show:

1. Blogs, community websites, and technical experience/reports are the most common GL sources used by SE researchers;
2. The main motivations to use GL is because its content could complement research findings by providing different results from scientific studies and answer practical and technical questions;
3. GL use is not widespread as a scientific reference due to some credibility and reliability constraints;
4. The use of the "shades of GL" can help SE researchers to assess GL and interpret the different GL types. Although, we identified that SE researchers have different interpretations of GL Control and Credibility;
5. The most relevant criteria used to assess GL credibility are the GL source be provided by renowned authors, institutions, companies, or be cited by a renowned source;
6. The most critical criteria to assess the Control and Credibility of a GL source are related to the producer reputation and the rigor of the GL content presented;
7. There is a positive correlation for credibility criteria considering the dimensions of Control and Credibility for each GL. It shows that when the level of Control increases, the level of Credibility tends to increase too;
8. We did not find significant differences between the opinions of graduate students and professors/researchers concerning the Control and Credibility dimensions analyzed of each GL type.

For replication purposes, all the data used in these investigations are available online at `https://doi.org/10.5281/zenodo.5164714`.

For future works, we plan i) To expand our view by investigating other SE research communities; and ii) To deeply understand the GL credibility aspects, focusing on building an objective quality assessment instrument that comprehends these several types.

References

Adams, J., Hillier-Brown, F. C., Moore, H. J., Lake, A. A., Araujo-Soares, V., and Summerbell, M. W. C. (2016a). Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. *Systematic Reviews*, 5(1):164.

Adams, R. J., Smart, P., and Huff, A. S. (2016b). Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, 19(4):432–454.

Baltes, S. and Ralph, P. (2021). Sampling in software engineering research: A critical review and guidelines.

Berg, V., Birkeland, J., Nguyen-Duc, A., Pappas, I. O., and Jaccheri, L. (2018). Software startup engineering: A systematic mapping study. *Journal of Systems and Software*, 144:255–274.

Bonato, S. (2018). *Searching the Grey Literature*. Rowman & Littlefield.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Coelho, J., Valente, M. T., Milen, L., and Silva, L. L. (2020). Is this GitHub project maintained? measuring the level of maintenance activity of open-source projects. *Information and Software Technology*, 1:1–35.

Dancey, C. P. and Reidy, J. (2004). *Statistics Without Maths for Psychology: Using Spss for Windows*. Prentice-Hall, Inc., USA.

Garousi, V., Felderer, M., and Hacaloğlu, T. (2017). Software test maturity assessment and test process improvement: A multivocal literature review. *Information and Software Technology*, 85:16–42.

Garousi, V., Felderer, M., and Mäntylä, M. V. (2016). The need for multivocal literature reviews in software engineering: Complementing systematic literature reviews with grey literature. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, EASE '16, pages 26:1–26:6, New York, NY, USA. ACM.

Garousi, V., Felderer, M., and Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivo-

cal literature reviews in software engineering. *Information and Software Technology*, 106:101–121.

Hosseinzadeh, S., Rauti, S., Laurén, S., Mäkelä, J.-M., Holvitie, J., Hyrynsalmi, S., and Leppänen, V. (2018). Diversification and obfuscation techniques for software security: A systematic literature review. *Information and Software Technology*, 104:72–93.

Kamei, F., Wiese, I., Lima, C., Polato, I., Nepomuceno, V., Ferreira, W., Ribeiro, M., Pena, C., Cartaxo, B., Pinto, G., and Soares, S. (2021). Grey literature in software engineering: A critical review. *Information and Software Technology*, page 106609.

Kamei, F., Wiese, I., Pinto, G., Ribeiro, M., and Soares, S. (2020). On the use of grey literature: A survey with the brazilian software engineering research community. In *Proceedings of the XXXIV Brazilian Symposium on Software Engineering*, SBES 2020, New York, NY, USA. Association for Computing Machinery.

Linåker, J., Sulaman, S., Maiani de Mello, R., and Martin, H. (2015). Guidelines for conducting surveys in software engineering. Technical report, Lund University.

Maro, S., Steghöfer, J.-P., and Staron, M. (2018). Software traceability in the automotive domain: Challenges and solutions. *Journal of Systems and Software*, 141:85 – 110.

Neto, G. T. G., Santos, W. B., Endo, P. T., and Fagundes, R. A. A. (2019). Multivocal literature reviews in software engineering: Preliminary findings from a tertiary study. In *Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '19, pages 1–6.

Oliveira, J. A., Viggiato, M., Pinheiro, D., and Figueiredo, E. (2021). Mining experts from source code analysis: An empirical evaluation. *Journal of Software Engineering Research and Development*, 9(1):1:1 – 1:16.

Petticrew, M. and Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*, volume 11. Blackwell Publishing Ltd.

Pinto, G., Ferreira, C., Souza, C., Steinmacher, I., and Meirelles, P. (2019). Training software engineers using open-source software: The students' perspective. In *Proceedings of IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training*, ICSE-SEET '19, pages 147–157. Institute of Electrical and Electronics Engineers (IEEE).

Rainer, A. (2017). Using argumentation theory to analyse software practitioners' feasible evidence, inference and belief. *Information and Software Technology*, 87:62–80.

Rainer, A. and Williams, A. (2018). Using blog articles in software engineering research: Benefits, challenges and case–survey method. In *Proceedings of the 25th Australasian Software Engineering Conference)*, ASWEC '18, pages 201–209.

Rainer, A. and Williams, A. (2019). Using blog-like documents to investigate software practice: Benefits, challenges, and research directions. *Journal of Software: Evolution and Process*, 31(11):e2197.

Raulamo-Jurvanen, Päivi, Mäntylä, M., and Garousi, V. (2017). Choosing the right test automation tool: A grey literature review of practitioner sources. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, EASE '17, pages 21–30. ACM.

Rodríguez-Pérez, G., Robles, G., and González-Barahona, J. M. (2018). Reproducibility and credibility in empirical software engineering: A case study based on a systematic literature review of the use of the szz algorithm. *Information and Software Technology*, 99:164–176.

Saltan, A. (2019). Do we know how to price saas: A multivocal literature review. In *Proceedings of the 2nd ACM SIGSOFT International Workshop on Software-Intensive Business: Start-Ups, Platforms, and Ecosystems*, IWSiB 2019, pages 7–12. ACM.

Schöpfel, J. and Prost, H. (2020). How scientific papers mention grey literature: a scientometric study based on scopus data. *Collection and Curation*.

Soldani, J., Tamburri, D. A., and Heuvel, W.-J. V. D. (2018). The pains and gains of microservices: A systematic grey literature review. *Journal of Systems and Software*, 146:215–232.

Spencer, D. (2009). *Card sorting: Designing usable categories*. Rosenfeld Media.

Storey, M.-A., Singer, L., Cleary, B., Filho, F. F., and Zagalsky, A. (2014). The (r) evolution of social media in software engineering. In *Proceedings of the on Future of Software Engineering*, FOSE '14. ACM Press.

Storey, M.-A., Zagalsky, A., Filho, F. F., Singer, L., and German, D. M. (2017). How social and communication channels shape and challenge a participatory culture in software development. *IEEE Transactions on Software Engineering*, 43(2):185–204.

Stray, V. and Moe, N. B. (2020). Understanding coordination in global software engineering: A mixed-methods study on the use of meetings and slack. *Journal of Systems and Software*, 170:110717.

Tom, E., Aurum, A., and Vidgen, R. (2013). An exploration of technical debt. *Journal of Systems and Software*, 86(6):1498–1516.

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.

Williams, A. (2018). Using reasoning markers to select the more rigorous software practitioners' online content when searching for grey literature. In *Proceedings of the 22Nd International Conference on Evaluation and Assessment in Software Engineering*, EASE '18, pages 46–56. ACM.

Williams, A. and Rainer, A. (2017). Toward the use of blog articles as a source of evidence for software engineering research. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, EASE'17, pages 280–285, New York, NY, USA. ACM.

Williams, A. and Rainer, A. (2019). How do empirical software engineering researchers assess the credibility of practitioner-generated blog posts? In *Proceedings of the 23nd International Conference on Evaluation and Assessment in Software Engineering*, EASE '19, pages 211–220. ACM.

Zahedi, M., Rajapakse, R. N., and Babar, M. A. (2020). Min-

ing questions asked about continuous software engineering: A case study of stack overflow. In Li, J., Jaccheri, L., Dingsøyr, T., and Chitchyan, R., editors, *EASE '20: Evaluation and Assessment in Software Engineering, Trondheim, Norway, April 15-17, 2020*, pages 41–50. ACM.

Zhang, H., Zhou, X., Huang, X., Huang, H., and Babar, M. A. (2020). An evidence-based inquiry into the use of grey literature in software engineering. In *Proceedings of the 42th International Conference on Software Engineering*, ICSE '20.