# About the Assessment of Grey Literature in Software Engineering

Guglielmo De Angelis
IASI-CNR
Roma, Italy
guglielmo.deangelis@iasi.cnr.it

Francesca Lonetti
ISTI-CNR
Pisa, Italy
francesca.lonetti@isti.cnr.it

## ABSTRACT

There is an ongoing interest in the Software Engineering field for multivocal literature reviews including grey literature. However, at the same time, the role of the grey literature is still controversial, and the benefits of its inclusion in systematic reviews are object of discussion. Some of these arguments concern the quality assessment methods for grey literature entries, which is often considered a challenging and critical task. On the one hand, apart from a few proposals, there is a lack of an acknowledged methodological support for the inclusion of Software Engineering grey literature in systematic surveys. On the other hand, the unstructured shape of the grey literature contents could lead to bias in the evaluation process impacting on the quality of the surveys. This work leverages an approach on fuzzy Likert scales, and it proposes a methodology for managing the explicit uncertainties emerging during the assessment of entries from the grey literature. The methodology also strengthens the adoption of consensus policies that take into account the individual confidence level expressed for each of the collected scores.

## CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; General literature; • **Software and its engineering**;

## KEYWORDS

Grey Literature, Quality Assessment, Likert scale, Fuzzy rating scale

**ACM Reference Format:**
Guglielmo De Angelis and Francesca Lonetti. 2021. About the Assessment of Grey Literature in Software Engineering. In *Evaluation and Assessment in Software Engineering (EASE 2021), June 21–23, 2021, Trondheim, Norway.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3463274.3463362

## 1 INTRODUCTION

Multivocal Literature Review (MLR) is a type of secondary study which takes into account blogs, videos, technical reports, white papers, and web-pages (i.e., the grey literature – GL) in addition to the published peer-review academic papers. MLRs aim to fill the gap between academic research and professional practice. While the importance of Systematic Literature Review (SLR) for documenting and reviewing different knowledge in the academic literature has been now fully recognised and accepted [10, 19], the role of the GL is controversial, then benefits and challenges of its inclusion in systematic reviews are object of discussion [1, 16].

Some meta-analysis guidelines recommend considering the items for the GL as long as their entries meet the inclusion/exclusion criteria [18]. However, searching for GL is challenging since differently from peer-reviewed literature: contents in GL are usually not collected or organised into libraries and databases; also often they miss of bibliographic information. Indeed, the time, the effort, and costs required in identifying, and retrieving the GL often make its inclusion prohibitive. Moreover, GL could be frequently incomplete, and its quality may be difficult to assess [18].

MLRs are traditionally popular in several fields (e.g., educational, social and medical sciences), but only recently they started to emerge as a type of secondary study in Software Engineering (SE). Recent works show indeed an ongoing interest in including GL in systematic reviews with the purpose to combine the academic state-of-the art and its practice in a field such as SE that deserves relevant attention to industrial concerns [8]. The GL allows to catch all the information that is constantly produced by SE practitioners outside of academic forum and can provide a valid feedback *from-the-filed* on both methodological and technological approaches.

Nevertheless, among the other issues preventing a widespread adoption of GL in SE research, there is the lack of both methodological support, and specific guidelines for GL inclusion in systematic surveys [25]. An attempt to fulfil this gap is represented by the Garousi et al.'s guidelines about how to include GL and conduct MLRs in SE [9].

In this paper we propose a methodology for assessing GL in SE leveraging both the well-known SLR guidelines by Kitchenham et al.'s [13], and the experience-based guidelines for MLR in [9]. Important and challenging aspects of the proposed methodology concern the quality assessment of GL entries, and the experiences matured in the field of Decision-Making Processes when dealing with subjective concepts.

SLRs consider only entries that undertook controlled peer-reviews and well-established publication process. In addition, there are available precise criteria on how to write and assess the different types of articles in the SE field [2, 21], thus the structures of candidate entries for SLR are somehow homogeneous. All these factors make the quality assessment process of Primary Studies in a SLR less dependent from the personal expertise and understanding of the reviewers. Concerning GL, the quality assessment process of the collected entries suffers of a more severe influence from the subjective evaluations of a reviewer. In other words, both the vagueness


*EASE 2021, June 21–23, 2021, Trondheim, Norway*

ACM ISBN 978-1-4503-9053-8/21/06…$15.00
https://doi.org/10.1145/3463274.3463362

and no-structured shape of the GL contents lead to biased evaluations [17] of the entries which introduce some *uncertainty* in the GL assessment process.

In order to mitigate potential subjective opinions, specific types of quality assessment checklists for the GL entries have been proposed [9, 24]. These checklists usually allow to express a judgement either by means of binary decision ("yes" or "no"), or by means of a 3-point Likert scale ("agree", "partly agree", and "disagree") [5]. The evaluation of all the items from the checklist leads to the decision to whether include or not a GL entry in the set of Primary Studies. An apparent advantage of approaches based on traditional Likert scales as the one proposed by Da Silva et al. [5] is that they aim to consider the *uncertainty* associated to the subjective judgement of the reviewer. However, as discussed in the following, the existing approaches for GL assessment do not allow to explicitly deal with the *magnitude of uncertainty* associated to each expressed judgement.

This work also aims to fill such a gap by combining both Likert scale and fuzzy rating scales for the quality assessment of GL entries. The proposed methodology leverages the fuzzy Likert scale approach proposed in [14] and it strengthens the importance of managing the *uncertainty* associated to the decision of including a GL entry during the review process. Specifically, it provides a more fine-grain estimation of the *uncertainty* level associated to the inclusion/exclusion of a GL entry: i) by enlarging the spectrum of possible agreement judgments that the reviewer can express (for instance 5 judgments are considered, i.e., "fully disagree", "partially disagree", "neutral", "partially agree" and "fully agree"); and ii) allowing the definition of a reviewer' confidence level on the expressed judgment.

However, our proposal to leverage the fuzzy Likert scale approach for the quality assessment of existing documents is very general and can be applied not only to GL entries but also for the evaluation of the qualitative studies that is influenced by the subjective judgement.

The rest of the paper is structured as follows: Section 2 presents some basic concepts about GL and its adoption in SE; Section 3 overviews existing fuzzy approaches for decision-making process; Section 4 shows the proposed methodology inside a quality assessment process and a reference instance as possible example; finally Section 5 presents conclusions and future research directions.

## 2 GREY LITERATURE

The most common definition of GL, i.e. the Luxembourg definition proposed at the Third International Conference on Grey Literature (ICGL), states that: "grey literature is produced on all levels of government, academics, business and industry in print and electronic formats, but it is not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body" [20]. Differently from academic publications, according to this definition, GL includes non-conventional documents, such as for instance, technical reports, blogs, video, theses and official documents that are easily available on internet.

The inclusion of GL in systematic reviews seems to follow an increasing trend over the time. In a survey of 1993, 77.7% of the meta-analysts and methodologists and only 47% of journal editors, thought that unpublished material should definitely or probably be included in scientific overviews [4]. The main concerns about the inclusion of GL were the lack of peer review and quality of the studies found in the grey literature. Another survey of 2006 showed that approximately 90% of systematic reviewers and approximately 70% of editors thought GL probably or definitely should be eligible for inclusion in systematic reviews [23].

The authors of [18] examine the effects of including GL in meta-analyses in fields that are different from SE, fostering the inclusion of all reports, grey and published, that meet predefined inclusion criteria.

The Grey Literature Network Service (GreyNet)[1] that is the organisation in charge of research, publication, open access, education, and public awareness of GL, promotes computer science among the top five subjects of GL according to the content of its databases in 2019. In particular, in the last years the use of GL is becoming widespread in SE research, where a huge amount of GL is produced and made available. Recent works focus on the challenges and benefits of the use of GL in SE [8, 25, 26]. Garousi et al. [8] outline the importance of performing Multivocal Literature Review (MLR) in a practitioner-oriented field such as SE. They also show what type of information is missed in some SE SLRs not including the grey literature sources and the advantages in terms of industrial needs that the SE community could have in performing MLR.

The authors of [25] carried out a systematic literature review on the use of GL in SE with the aim of empirically investigating the SE researchers'views and outlining the challenges as well as possible solutions about the use of GL in SE. This study evidences the potentiality of GL of becoming a valid source of information able to complement the white literature in SE research.

Finally the authors of [26] identify 102 secondary studies in SE published by June 2019 that include GL. By this study, five main factors motivate the inclusion of GL that are: i) looking for more related results; ii) avoiding publication bias; iii) comparing different perspectives between researchers and practitioners; iv) understanding the views of the practitioner's community; and finally v) exploring uncharted research areas.

Although the existing SLR guidelines by Kitchenham et al.'s [13] briefly hint the idea of including GL sources in SLR studies, they do not provide precise guidelines for how to treat GL. To fill this gap, Garousi et al. [9] provide a methodology as well as experience-based guidelines for planning, conducting and presenting MLR studies in SE. They analyse a set of 24 MLR guidelines and experience papers in other fields and taking as reference the SLR guidelines by Kitchenham et al.'s [13] they focus on the steps that are different for conducting MLRs, explaining them through a MLR running example. They also provide a set of criteria that should be met to decide whether to include the GL in a review study (then conduct an MLR study) or perform a conventional SLR.

In this paper, we refer to existing guidelines for SLR [13] and MLR [9] focusing on the quality assessment process of the GL entries, and we propose a fuzzy Likert scale based methodology to cope with the *uncertainty* level associated to the inclusion/exclusion of GL entries. We describe in detail this methodology in Section 4.

[1] http://www.greynet.org/

## 3 FUZZY APPROACHES IN DECISION-MAKING PROCESSES

Decision-Making Processes concern a set of activities in which several decision makers cooperate analysing and evaluating the subjects of the decision, often giving indications for the selection among several alternatives. Usually, these processes rely on statistical data that have been recorded from some kind of empirical experiment (e.g. surveys, questionnaire studies, scientific inquiries). Also, decision makers are often requested to interact in order to achieve some level of agreement or consensus on the decisions about the observed subjects [3].

In any kind of empirical experiment, the observed variables can be either measured using an objective measurement system, or estimated from Human being perception. Whenever intrinsically subjective concepts have to be assessed (e.g., relevance, difficulty, perceived workload, feelings), subjective variables are present [15]. In these cases, the referred measurement method includes both set of answers by individuals (either an expert or not), and a framework allowing decision makers to give their assessment.

It is important to remark that ambiguities are usually unavoidable when dealing with subjective concepts, thus *uncertainty* becomes a common factor in a wide range of real-world decision-making problems [17]. This uncertainty could be due to the vagueness of the information reported by the individuals involved in the empirical experiment, or due to a misleading understanding on the meaning of the adopted terms.

Within the context of empirical experiments adopting questionnaires, the well-known Likert scale is one of the most commonly referred frameworks in order to collect opinions/judgements among admissible options [6]. For example, most of the evaluation methods in the educational context employ Likert scales [12]. Possibly, the popularity of this scale is because of it facilitates the survey construction, data collection, and analysis [14]. The Likert scale is built on a set of ordinal linguistic variables encoded by means of integer numbers [11]. Some specific studies in the literature argue that adopting five-point scale helps in reducing the "laziness" effect when answering the questionnaire [11]. Nevertheless, matching linguistic variables with integer numbers is considered a complex task as the alternatives may be not equally important for the respondents, and the differences between the expressed options cannot be interpreted in terms of their magnitude [12].

The fuzzy measurement systems have been proved as a valid alternative to *crisp approaches* [15] (e.g., traditional Likert scales) especially when dealing with decisions taken assessing questionnaires on subjective concepts [6]. Indeed, respondents to questionnaires may not always have a clear judgement about where a subject fits in a set of admissible options [14].

Instead of assigning an exact answer, fuzzy systems either allow to express opinions/judgements on different granularity of uncertainty they want to manage, or to identify the range of possible scores that reflect a respondent's confidence in the given answer. Specifically, *fuzzy linguistic approaches* encode uncertainty by means of linguistic descriptors, which are implicitly assumed to match with fuzzy numbers or intervals [6]. The matching between linguistic descriptors and fuzzy intervals is frequently considered as an a-posteriori activity done in order to encode the collected data in a blurred space [7]. Fuzzy linguistic approaches are expected to assess the degree of consensus among decision makers in a more flexible way than crisp approaches, reflecting the large spectrum of possible agreements and guiding the discussion process until widespread (or even partial) agreement is achieved among the group of decision makers [3].

However, fuzzy linguistic approaches may present few limitations. Among the others, the modelling of the subjective information by means of linguistic terms may mislead with the personal understanding a respondent has about the semantic of the referred concrete terms [17]. Approaches based on the *fuzzy rating scale* mitigate such limitations by combining both the positive features from fuzzy linguistic scales and the possibility to relate a fuzzy confidence to the expressed opinion. In this sense, such approaches explicitly allow respondents to cope with imprecision while expressing a judgement. Fuzzy rating scales are also considered as a-priori means for assessing the continuous nature of subjective interpretations into fuzzy values [7].

## 4 GREY LITERATURE QUALITY ASSESSMENT

In this section we describe the proposed methodology based on a Fuzzy Likert scale for the quality assessment of GL entries. We first present the general methodology in Section 4.1, and then a theoretical example of its instantiation in Section 4.2.

### 4.1 The Methodology

This section presents an overview of the proposed methodology for assessing GL entries. The description is also depicted in Figure 1 as UML Activity Diagram. The proposed methodology refers to the phases of planning and conducting SLR of Kitchenham et al.'s guidelines [13] and it leverages the Fuzzy Likert approach proposed in [14]. Specifically, in Figure 1 we depict in dark grey color the new activities introduced in our methodology that are not included in the Kitchenham et al.'s guidelines. Whereas, in light grey color we depict the activities that are foreseen in the guidelines in [13] but adapted in order to deal with the proposed fuzzy quality assessment.

The initial set of activities in Figure 1 concerns the planning of the GL review. Specifically, the structured activity "Plan the Review of the Gray Literature" refers to the identification of the GL sources (i.e. the archives where to look for GL entries, or the engines for retrieving them), also the search query as well as the inclusion and exclusion criteria are defined. For more details of this activity we refer to the existing SLR guidelines [13].

The structured activity "Plan the Quality Assessment" in Figure 1 focuses on the quality assessment planning. Specifically, it includes: **i)** the definition of a set of indicators as a checklist. An example of quality assessment checklist has been developed in [9]; whereas, Table 1 reports the set of quality indicators later adopted in Section 4.2; **ii)** the identification of a set of reviewers for processing the entries. Reviewers may have different roles. For instance, in Section 4.2 three reviewers are identified, but only two of them are assigned to the quality assessment of each entry, whereas the third one is in charge of enforcing the policies on consensus as described below; **iii)** the definition of the Fuzzy Likert scale derived by the combination of a Likert scale and a Fuzzy rating scale according to [14]. Differently from a Likert scale, the proposed Fuzzy Likert

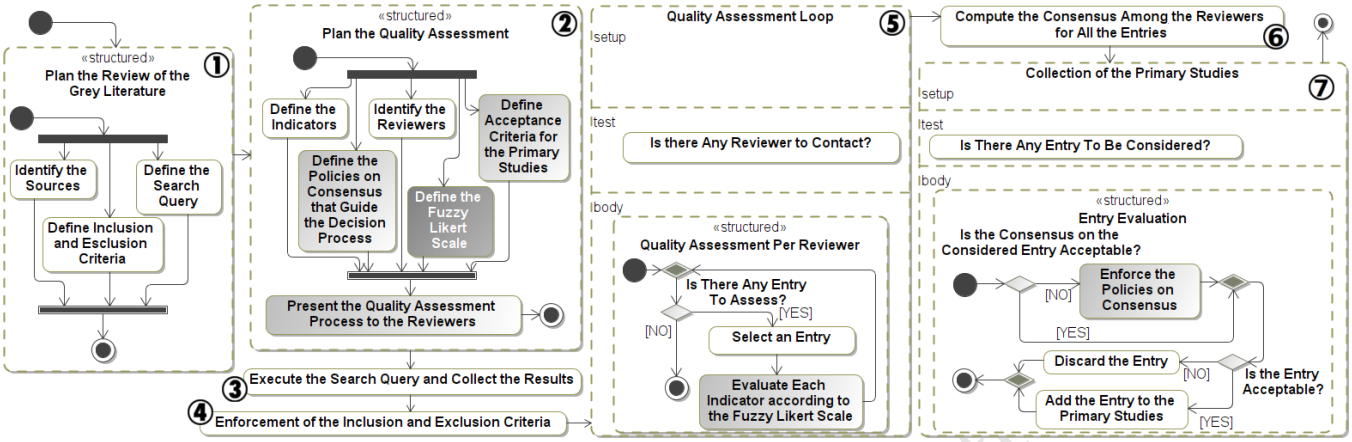


**Figure 1: Proposed Methodology for Assessing GLs**

scale allows to explicitly manage the *uncertainty* level intrinsically associated to the quality assessment of GL entries. An example of the fuzzification and de-fuzzification applied to each quality indicator is showed in Section 4.2. As well, the example reports about the aggregation of the set of de-fuzzified quality scores into an overall quality score per reviewer; **iv)** the definition of acceptance criteria for the Primary Studies. In Section 4.2 the overall quality score of an entry per review is acceptable when it is higher than 3; **v)** the definition of the policies on the consensus that will guide the decision process. For instance, in Section 4.2 the policy on the consensus states that when the agreement between the two reviewers is less than 0.85, the entry is processed by a third review; **vi)** finally, the planned quality assessment process is presented to the reviewers.

The remaining five activities of our methodology are related to conducting the review. In particular, in the third activity the search query is executed on the identified sources and a set of entries is collected to be analysed; then in the fourth activity the inclusion and exclusion criteria are applied to these entries. For details about the third and fourth activities we refer to the existing SLR [13] and MLR [9] guidelines.

The resulting set of entries is considered for the quality assessment (i.e. the activity "Quality Assessment Loop"). Specifically, during this activity each entry is processed, possibly by several reviewers. Each reviewer expresses his/her opinion on all the indicators defined during the planning of the quality assessment according to the defined fuzzy Likert scale. For each indicator, the reviewer selects two consecutive quality scores (each one respectively associated to a judgment, see Table 2) and their respective confidence. Then, a de-fuzzified quality score is computed for each indicator. The overall quality score for each reviewer is computed by considering the de-fuzzified Likert values associated to all the indicators. For instance, the overall quality score in Section 4.2 is computed as the average on the de-fuzzified Likert values from the 8 indicators in Table 1. In this way, every entry results associated with as many overall quality scores as the number of the reviewers processed it. In other words, each overall quality score encodes the review opinion on the quality of an entry.

Next the consensus among the reviewers is estimated starting from their overall quality scores. An example of consensus model among two reviewers is showed in Equation 1.

The last structured activity focuses on the definition of the set of Primary Studies (i.e. the loop activity "Collection of the Primary Studies"). Specifically, for each entry to be considered, if the level of consensus among the reviewers is not acceptable, then the consensus policies are enforced. For instance, such an enforcement could foresee to run a meeting among the reviewers who performed the quality assessment, or to rely on the judgement of another reviewer. If the consensus among the reviewers results acceptable, the quality of the entry is evaluated against the acceptance criteria in order to assess whether it has to be included or not among the Primary Studies.

## 4.2 An Instantiation

In this section we provide an instance of the proposed methodology and we discuss in detail how some of its activities are applied to a theoretical example (i.e., the activities named "Plan the Quality Assessment", "Quality Assessment Loop", "Compute the Consensus Among the Reviewers for All the Entries" and "Collection of the Primary Studies").

During the planning of the quality assessment (i.e activity "Plan the Quality Assessment" in Figure 1) we assume to have: **i)** a checklist aggregating 8 different indicators as reported in Table 1; **ii)** 3 reviewers in charge of performing the quality assessment; **iii)** an accepting criterion stating that the overall quality score of each entry must be greater than 3; **iv)** the definition of the Fuzzy Likert scale as described in Table 2; **v)** the definition of the consensus model.

In detail, the Quality Assessment plan foresees that each reviewer for each entry selects two *consecutive*[2] quality scores expressing a judgement with respect to the considered indicator (see Table 2). Also, she/he has to give a confidence level for each of these two quality scores. The sum of the two confidence levels expressed per indicator is assumed to be 1.

[2]Consecutive in the Fuzzy Likert Scale in Table 2.

| | |
|---|---|
| I1 | The publishing organization is supposed to be authoritative |
| I2 | The authors are associated with a renowned organization |
| I3 | The authors published other work in the field |
| I4 | The authors have a clear expertise in the field |
| I5 | The document is focusing on the area of interest |
| I6 | The source statements are as much objective as possible |
| I7 | The source has a clearly stated aim |
| I8 | The source has a stated methodology |

**Table 1: Indicators driving the quality score procedure**

| Judgements | Fully Disagree | Partially Disagree | Neutral | Partially Agree | Fully Agree |
|---|---|---|---|---|---|
| Scores | 1 | 2 | 3 | 4 | 5 |

**Table 2: Fuzzy Quality Judgements and Scores**

For each fuzzy quality score expressed by a reviewer (i.e., the two pairs of quality scores and their confidence), the Quality Assessment plan relies on the de-fuzzification process described in [14]. A de-fuzzification process represents the transformation procedure that maps back a fuzzy input into a scalar value. More specifically, in [14] the de-fuzzification process has been instantiated by taking into account the triangular isosceles membership functions; the same approach is referred in this methodology instance within the activity "Plan the Quality Assessment".

This same activity foresees that the overall opinion each reviewer has about an entry will be calculated by taking into account all the quality scores from the 8 indicators. Specifically, as all the indicators have the same relevance for this example, thus the overall entry's quality score per reviewer will be computed as the average on the resulting de-fuzzified Likert values.

The Quality Assessment plan establishes the consensus as a measure on the overall quality scores from all the reviewers [22]. Intuitively, consensus is modelled as a function over a set of different opinions about some statements expressed on the basis of a pre-defined scale (e.g., Likert scale, or fuzzy Likert scale) and that ranges from 0 (i.e. complete disagreement of opinions), to 1 (i.e., complete agreement). Following the formulation reported in [22], the consensus model adopted in this example is given with Eq. 1

$$Cns(\begin{bmatrix} r_1 \\ r_2 \end{bmatrix}) = 1 + \sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{\mid r_i - \overline{R} \mid}{d} = 1 + \log_2 \frac{\mid r_1 - \overline{R} \mid}{d} = \\ = 1 + \log_2 \frac{\mid r_2 - \overline{R} \mid}{d} \quad (1)$$

where $r_i$ is the overall quality score from the reviewer $i$ on the considered entry; $d = L_{max} - L_{min} = 4$ is the width of categories on the referred Fuzzy Likert scale (i.e., $5 - 1 = 4$), and $\overline{R}$ is the mean on the overall quality scores by the two reviewers.

During the quality assessment (i.e. the "Quality Assessment Loop" activity of Figure 1), all the entries are processed by two reviewers who assign to each indicator two quality judgements and related confidence levels as described above.

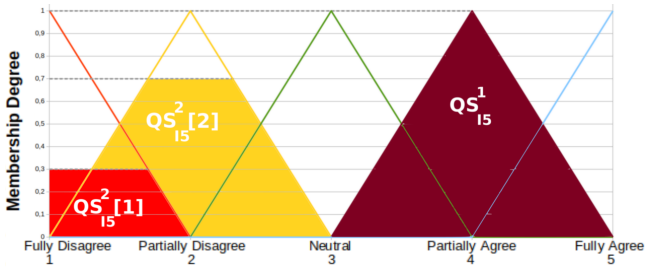


**Figure 2: Triangular Isosceles Membership Functions**

For instance, let us consider the indicator I5: a reviewer can have a clear opinion about the focus of the contribution (e.g., she/he somehow agrees with the indicator), or she/he tends to disagree with the statement associated to the indicator but without a crystal opinion. The former case could be represented by assigning a single judgement (e.g., $QS^1_{I5} = \{[4; 1.0]\}$), while in the latter the reviewer can express the fuzzy judgement by means of a pair of quality scores with different confidence degrees (e.g., $QS^2_{I5} = \{[1; 0.3], [2; 0.7]\}$).

According to the triangular isosceles membership function adopted during the activity "Plan the Quality Assessment", the output of the de-fuzzification process is calculated as a combination of the quality scores in the two fuzzy pairs (i.e., $J_i$ and $J_{i+1}$ in Eq. 2) weighted with the area of the trapezoids resulting from their respective confidence degrees (i.e., $\mathcal{A}(C_i)$ and $\mathcal{A}(C_{i+1})$ in Eq. 2).

$$Output = \frac{J_i \mathcal{A}(C_i) + J_{i+1} \mathcal{A}(C_{i+1})}{\mathcal{A}(C_i) + \mathcal{A}(C_{i+1})} \quad (2)$$

Figure 2 depicts the triangular isosceles membership functions; while the values 4, and 1.78 are the results of the de-fuzzification process when applied to the respective fuzzy quality scores $QS^1_{I5}$, and $QS^2_{I5}$ from the example above.

The decision about adding or rejecting an entry from the set of Primary Studies is taken by estimating the level of agreement reached by both the reviewers (i.e. the activity "Compute the Consensus Among the Reviewers for All the Entries" in Figure 1) according to the consensus model defined within the activity "Plan the Quality Assessment", and discussed above.

Referring to the last activity of Figure 1, i.e. "Collection of the Primary Studies", on the one hand, when the two reviewers show an high agreement on a given entry (i.e., the consensus rates at least 0.85, see Table 3) then the verdict on their evaluations is assumed to be significant. In this case, if the average on both the overall quality scores is greater than 3 the entry is added to the set of the Primary Studies; otherwise it is rejected. On the other hand, when the agreement between the two reviewers is not significant (i.e. the consensus rates less than 0.85, see Table 3) then the entry is processed by a third reviewer who decides if it deserves to be included or not within the set of Primary Studies.

## 5 CONCLUSIONS AND FUTURE WORK

The interest on including GL in systematic surveys is growing in the last years in SE field. The evaluation of GL entries is challenging due to both their unstructured shape, and often the poor organization of their content. The lack of established criteria for GL assessment

| | | Average of the scores by 2 reviewers | |
|---|---|---|---|
| | | <= 3 | > 3 |
| Consensus | < 0.85 | $3^{rd}$ reviewer decides | $3^{rd}$ reviewer decides |
| | >= 0.85 | Excluded | Included |

**Table 3: Acceptance Criteria Driven by the Consensus**

impacts on the quality evaluation of GL entries; thus the assessment process becomes more conditioned by the subjective judgement of the reviewers.

The existing guidelines for performing GL studies in SE provide limited coverage about the methods for addressing the *uncertainty* associated to the subjective judgement of the reviewer. In particular, they do not allow to explicitly deal with the *magnitude of uncertainty* associated to each expressed judgement.

This paper filled this gap by providing a methodology for quality assessment of the GL entries leveraging a fuzzy Likert scale. To better manage the *uncertainty* associated to the reviewers judgments, the proposed methodology allows the reviewers to express a set of possible agreement judgmnets as well as a confidence level on the expressed judgment. The activities of the proposed methodology have been described referring to the phases of planning and conducting SLR of Kitchenham et al.'s guidelines [13]. As an example, a reference instance of the proposed methodology has been also presented.

In the future, we plan to apply the proposed methodology for conducting GL reviews on different topics of SE. We want to improve the proposed methodology on the base of the experience and lesson learned of applying it. We also aim to refine the proposed methodology according to the different types of GL sources or the specific SE areas addressed by the GL. As long term research we would like also to provide some evidence that using a fuzzy Likert approach has some benefits with respect to the adoption of regular Likert scales in the assessment of GL entries.

Finally, another future goal is to include the proposed methodology inside a more comprehensive set of guidelines about how to perform the GL review in SE, and share these guidelines within the community.

## ACKNOWLEDGMENTS

This paper has been supported by the Italian MIUR PRIN 2017 Project: SISMA (Contract 201752ENYB), and partially by the Italian Research Group: INdAM-GNCS.

## REFERENCES

[1] Jean Adams, Frances C Hillier-Brown, Helen J Moore, Amelia A Lake, Vera Araujo-Soares, Martin White, and Carolyn Summerbell. 2016. Searching and synthesising 'grey literature'and 'grey information'in public health: critical reflections on three case studies. *Systematic reviews* 5, 1 (2016), 1–11.

[2] Antonia Bertolino, Antonello Calabrò, Francesca Lonetti, Eda Marchetti, and Breno Miranda. 2018. A categorization scheme for software engineering conference papers and its application. *J. Syst. Softw.* 137 (2018), 114–129.

[3] Francisco Javier Cabrerizo, Francisco Chiclana, Rami Al-Hmouz, Ali Morfeq, Abdullah Saeed Balamash, and Enrique Herrera-Viedma. 2015. Fuzzy decision making and consensus: challenges. *Journal of Intelligent & Fuzzy Systems* 29, 3 (2015), 1109–1118.

[4] Deborah J Cook, Gordon H Guyatt, Gerard Ryan, Joanne Clifton, Lisa Buckingham, Andrew Willan, William Mcllroy, and Andrew D Oxman. 1993. Should unpublished data be included in meta-analyses?: Current convictions and controversies. *Jama* 269, 21 (1993), 2749–2753.

[5] Fabio QB Da Silva, André LM Santos, Sérgio Soares, A César C França, Cleviton VF Monteiro, and Felipe Farias Maciel. 2011. Six years of systematic literature reviews in software engineering: An updated tertiary study. *Information and Software Technology* 53, 9 (2011), 899–913.

[6] Sara de la Rosa de Sáa, María Ángeles Gil, Gil González-Rodríguez, María Teresa López, and María Asunción Lubiano. 2014. Fuzzy rating scale-based questionnaires and their statistical analysis. *IEEE Transactions on Fuzzy Systems* 23, 1 (2014), 111–126.

[7] Pierpaolo D'Urso. 2017. Exploratory multivariate analysis for empirical information affected by uncertainty and modeled in a fuzzy manner: a review. *Granular Computing* 2, 4 (2017), 225–247.

[8] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 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)*. Article 26, 6 pages.

[9] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information & Software Technology* 106 (2019), 101–121.

[10] S Gopalakrishnan and P Ganeshkumar. 2013. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of family medicine and primary care* 2, 1 (2013), 9.

[11] James Hartley. 2014. Some thoughts on Likert-type scales. *International journal of clinical and health psychology* 14, 1 (2014), 83–86.

[12] Tamás Jónás, Zsuzsanna Eszter Tóth, and Gábor Árva. 2018. Applying a fuzzy questionnaire in a peer review process. *Total Quality Management & Business Excellence* 29, 9-10 (2018), 1228–1245.

[13] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).

[14] Qing Li. 2013. A novel Likert scale based on fuzzy sets theory. *Expert Systems with Applications* 40, 5 (2013), 1609 – 1618.

[15] Pierre Loslever, Taisa Guidini Gonçalves, Káthia Marçal de Oliveira, and Christophe Kolski. 2019. Using fuzzy coding with qualitative data: example with subjective data in human-computer interaction. *Theoretical Issues in Ergonomics Science* 20, 4 (2019), 459–488.

[16] Quenby Mahood, Dwayne Van Eerd, and Emma Irvin. 2014. Searching for grey literature for systematic reviews: challenges and benefits. *Research synthesis methods* 5, 3 (2014), 221–234.

[17] Sebastia Massanet, Juan Vicente Riera, Joan Torrens, and Enrique Herrera-Viedma. 2014. A new linguistic computational model based on discrete fuzzy numbers for computing with words. *Information Sciences* 258 (2014), 277–290.

[18] Laura McAuley, Peter Tugwell, David Moher, et al. 2000. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *The Lancet* 356, 9237 (2000), 1228–1231.

[19] Ralf W Schlosser. 2006. The role of systematic reviews in evidence-based practice, research, and development. *Focus* 15 (2006), 1–4.

[20] Joachim Schöpfel and Dominic J Farace. 2010. Grey literature. In *Encyclopedia of library and information sciences*. 2029–2039.

[21] M. Shaw. 2003. Writing good software engineering research papers. In *Proceedings of 25th International Conference on Software Engineering*. 726–736.

[22] William J. Tastle and Mark J. Wierman. 2007. Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning* 45, 3 (2007), 531 – 545.

[23] Jennifer Tetzlaff, David Moher, Ba Pham, and Douglas Altman. 2006. Survey of views on including grey literature in systematic reviews. In *16th Cochrane Colloquium, Dublin, Ireland*.

[24] J Tyndall and J Tyndall. 2010. AACODS checklist. *Flinders University* (2010).

[25] H. Zhang, X. Zhou, X. Huang, H. Huang, and M. A. Babar. 2020. An Evidence-Based Inquiry into the Use of Grey Literature in Software Engineering. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*. 1422–1434.

[26] Xin Zhou. 2020. How to Treat the Use of Grey Literature in Software Engineering. In *Proceedings of the International Conference on Software and System Processes (ICSSP '20)*. 189–192.