Noname manuscript No.
(will be inserted by the editor)

Reasoning about Research Quality Alignment in Software Engineering

Jefferson Seide Molléri · Michael Felderer · Kai Petersen · Emilia Mendes

Received: date / Accepted: date

Abstract Context: Research quality is intended to assess the design and reporting of studies. It comprises a series of concepts such as methodological rigor, practical relevance, and conformance to ethical standards. Depending on the perspective, different views of importance are given to the conceptual dimensions of research quality.

Objective: We aim to better understand what constitutes research quality from the perspective of the empirical software engineering community. In particular, we intend to assess the level of alignment between researchers with regard to a conceptual model of research quality.

Method: We conducted a mixed methods approach comprising an internal case study and a complementary focus group. We carried out a hierarchical voting prioritization based on the conceptual model to collect relative values for importance. In the focus group, we also moderate discussions with experts to address potential misalignment.

Results: We provide levels of alignment with regard to the importance of quality dimensions in the view of the participants. Moreover, the conceptual model fairly expresses the quality of research but has limitations with regards the structure and description of its components.

Conclusion: Based on the results, we revised the conceptual model and provided an updated version adjusted to the context of empirical software engineering research. We also discussed how to assess quality alignment in research using our approach, and how to use the revised model of quality to characterize an assessment instrument.

Keywords Research Quality · Alignment · Mixed Method · Case Study · Focus Group

Jefferson Seide Molléri
Tel.: +046 0734223608
E-mail: jefferson.molleri@bth.se

1 Introduction

Research quality is a central concept in software engineering research. In the context of evidence-based software engineering, the quality plays an important role when aggregating evidence. Quality standards are a means for evaluating the research quality. Quality standards are a means for evaluating the research quality, and are often operationalized as checklists to assess quality (e.g., [1,2,3]).

Despite having specific standards for different research methods, generic standards are of importance for the quality of research as a whole. Thus, Mårtensson et al. [4] proposed a generic framework of standards that research should exhibit. The main dimensions are: credible, contributory, communicable, and conforming. Within the rigorous category, we find subcriteria related to validity and methodological stringency. The contributory category focuses on the originality and relevance of the research. Communicable refers to the presentation of the research, and conforming is mainly concerned with ethical aspects.

As is the case with the quality of software not all qualities of research can be achieved to the same degree at once. For example, a system that should be highly secure may not have a high degree of usability due to authentication mechanisms a user may have to go through. Hence, software engineers need to make a trade-off between the quality dimensions by prioritizing them. Barney and Wohlin [5] investigated the priorities of quality standards among different groups to understand how different groups and roles differ.

Similarly, Siegmund et al. [6] investigated how the view of researchers differ with respect to two quality dimensions, namely the external and internal validity of studies. As an example for trade-offs they highlighted: *There is an inherent trade-off in empirical research: Do we want observations that we can fully explain, but with limited generalizability, or do we want results that apply to a variety of circumstances, but where we cannot reliably explain underlying factors and relationships? Due to the options' different objectives, we cannot choose both.*

Hence, Siegmund et al. [6] elicited preferences within the community in a very simple manner, namely by asking whether there is a preference for maximizing or rather minimizing internal or external validity. A key finding was that opinions varied greatly among the participants. Siegmund et al. did not discover other studies focusing on eliciting trade-off preferences. Hence, research on understanding preferences concerning research quality is needed as this informs which methods are chosen and how studies are designed.

In our study, we investigated the preferences of empirical software engineering (ESE) researchers concerning the importance of quality dimensions proposed by Mårtensson et al. [4]. Specifically, we make the following scientific contributions:

- We apply a method proposed by Barney and Wohlin [5] to determine the level of alignment among researchers on the importance given to dimensions of quality. This is interesting for research groups and communities to discuss their preferences and make the need for trade-offs explicit;
- We revise the conceptual research quality assessment model by Mårtensson et al. [4] and provide needs for adjusting to the context of empirical software engineering research. This is an important step to adapt the conceptual model to the view of potential community or the field it is intended to be applied.

– We propose an exemplary scenario to apply the revised model to characterizing a quality assessment instrument. This provides the potential users with insights on how to operationalize the model in a more realistic scenario.

While Siegmund et al. [6] focused on a larger part of the research community, we focused on two specific cases. The first was the internal case of the Software Engineering Research and Education Lab Sweden (SERL Sweden). The second was the community's case of the International Software Engineering Research Network (ISERN).

The remainder of the paper is structured as follows: Section 2 presents the related work. Sections 3 and 4 describe the research method and provides the results for the case study and focus group, respectively. We discuss the results in Section 5. Section 6 concludes the paper.

2 Background & Related Work

Evaluating research quality is a fundamental topic that has been discussed by researchers for a long time [7]. Such discussions lead to results that are formulated in terms of quality concepts aimed to assess research's methodological aspects. Examples of such concepts are reliability [8], practical relevance [9] and generalizability [10].

Furthermore, quality concepts are often addressed with regard to a particular methodology: validity and reliability are strongly related to a positivist approach, whereas the practical relevance is emphasized by interpretivist studies [11]. Some researchers argue on trade-offs needed, but others argue in favour of an integrated view, in which both concepts are relevant.

2.1 A Conceptual Model of Research Quality

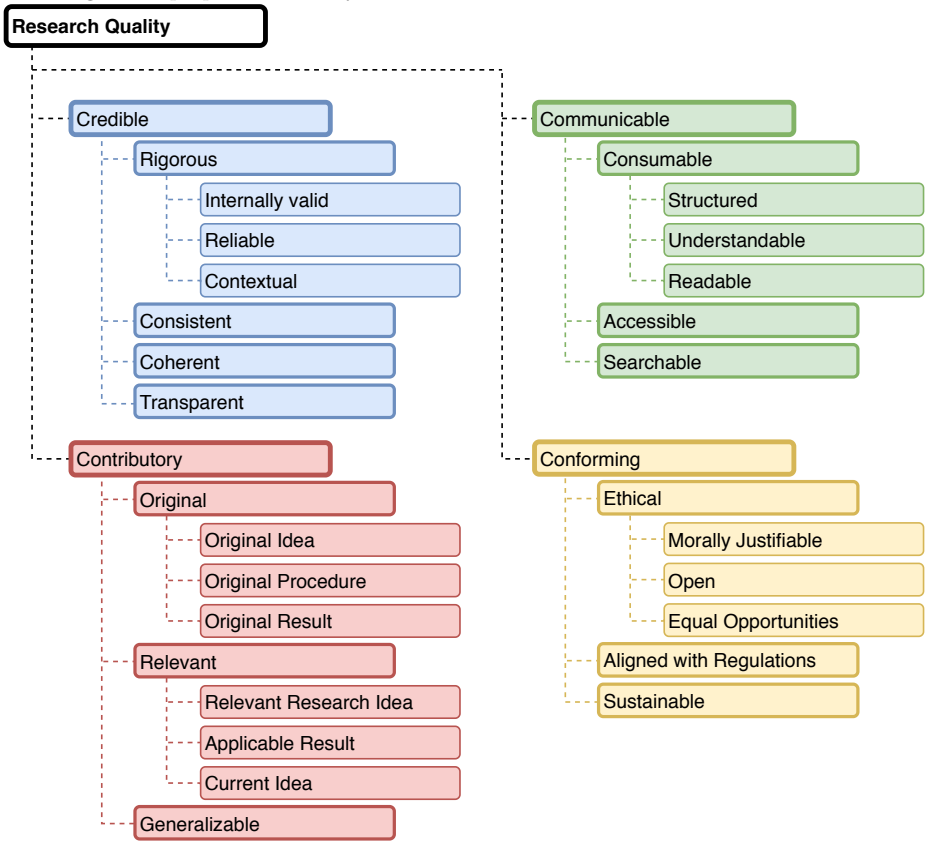
Aligned with an integrated view, Mårtensson et al. [4] proposed a generic framework for classification of research quality, comprising 23 dimensions divided into four main areas: A) credible, B) contributory, C) communicable, and D) conforming. Details of each criterion are given in Appendix A.1.

This framework is modelled as a hierarchical structure, in which related dimensions are grouped into subareas. For example, rigorous is defined as research that is contextual, internally valid and reliable. Figure 1 illustrates the hierarchical relationship between the dimensions, their main areas, and subareas.

Further, this framework aims to represent research quality regardless of research field; however, its authors acknowledge the value in assessing its suitability within the context of different research fields. Motivated by such, the work described herein examines the use of such framework in the context of the research community we belong to - Empirical Software Engineering community.

We aim to identify which dimensions are considered more important for this community, and also within sub-groups in the same community. Explicitly, we ought to determine the level of alignment between researchers regarding the perceived importance of the quality dimensions.

Fig. 1 A representation of the 23 dimensions proposed by Mårtensson et al. [4] organized according to the proposed hierarchy.



2.2 Methods to Determine Alignment

We used two methods to identify the level of alignment between researchers. First, the Stakeholder Alignment Assessment Method for Software Quality (SAAMSQ) [12] is a stepwise approach to elicit the importance assigned to different dimensions of a given quality model. It allowed us to determine which dimensions are subject to shared or conflicting views of importance. The method comprises 7 steps:

- Select a case context;
- Identify key stakeholders;
- Select/tailor a quality model;
- Develop a data collection instrument;
- Conduct data collection;
- Analyze the results; and
- Present the results to participants.

Although focused on the quality of software products and services, SAAMSQ has the potential to identify views of quality in different contexts. Also, we rec-

ognize similarities between software quality standards described in [13] and the conceptual model in [4] that supports the application in the context of our study.

Secondly, focus group allowed us to closely interact with researchers, fostering a discussion on the perceived importance of the quality dimensions. Focus group research employs mediated discussions to gather qualitative data from a group of participants regarding the investigated topic [24]. In general, focus group produces an in-depth understanding of the investigated phenomenon, its causes and implications.

We used the methods complementary to each other. SAAMSQ was first applied in the internal case of SERL Sweden (Section 3), in which we were more familiar to the participants. Insights gathered from this first study led us to design a more thorough focus group (Section 4) to both determine the alignment and discuss the opinions of the participants regarding the importance of quality dimensions.

3 Internal Case Study

3.1 Method

Our research methodology is a single case study employing a survey questionnaire to collect quantitative data. Specifically, the participants assigned values of importance to dimensions describing different aspects of research quality. We followed the guidelines by Runeson and Höst for conducting and reporting case study research [14].

3.1.1 Context

This study was conducted in the Software Engineering Research and Education Lab Sweden (SERL Sweden), part of the Department of Software Engineering in the Blekinge Institute of Technology (BTH)[1]. Currently, close to 50 people work at SERL, representing 18 different nationalities. A large part of the team consists of research practitioners, i.e. 7 professors, 2 associate professor, 13 assistant professor, and 19 PhD students. At the time we conducted our study the team was a little different, which can present discrepancy with the results provided in Section 3.2.1. As an example, Postdoctoral researchers were promoted to assistant professors.

In line with BTH's mission, SERL's research is grounded in innovation and close collaboration with industry. It has partnership with several companies that develop software intense systems, services, and products. Moreover, BTH is well-ranked academic research in empirical and evidence-based software engineering [15].

The members of SERL Sweden are also acknowledged for providing methodological guidelines for researchers, covering a wide range of methods such as experiments (e.g. [16]), case studies (e.g. [17]), systematic literature and mapping studies (e.g. [18,19]), and mixed methods studies (e.g. [20]).

[1] `https://www.bth.se/eng/about-bth/organisation/faculty-of-computing/dipt/`

Research questions and proposition

This case study's goal was to investigate the relative importance of the quality dimensions, as per the previously presented framework, based upon participants' judgements. Two research questions were formulated as means towards achieving such goal, as follows:

RQ1a What are the most important research quality dimensions in SE research, as per the aforementioned framework, according to the participants?

RQ1b How aligned are the different dimensions prioritizations, as per the different participants' assessments?

Our proposition for this research is that an alignment of the perceived importance of research quality represents the level of agreement within the context of this group. To recall, the study by Mårtensson [4] found very different preferences, which may originate from the spread of participants between different program committees with different foci.

3.1.2 Subjects

Subjects were selected through convenience sampling from the researcher positions at SERL, i.e. full professors, associate professors, assistant professors, and Ph.D. students. We sent invitations to the candidates and clustered the ones willing to participate in groups up to five.

3.1.3 Unit of analysis

The unit of analysis is the relative importance assigned to each dimension of research quality. Relative importance may show distances between the interpretations and permits to demonstrate which are the most critical dimensions in comparison to others.

We designed a data collection instrument to allow for subjects to express to what extent the dimensions are important. Subjects assigned importance values using hierarchical cumulative voting [21]. First, they should distribute 1000 points between the four main areas (Credible, Contributory, Communicable, and Conforming) in the Framework, and later distribute another 1000 between the subcriteria related to each area. In this way, the values for each subcriteria are weighted according to the values assigned to its area. Here the higher the number of points, the higher the importance (and hence priority).

Cumulative voting encourages subjects to prioritize and make trade-offs, given that whenever they add points to one criterion, they have to also remove points from another criterion, so to keep the total intact. This way it becomes more difficult to simply give the same number of points to every criterion.

3.1.4 Data collection

We conducted four workshops to gather data from 17 participants regarding the importance of the Framework's quality dimensions. Due to convenience, all workshops took place in a meeting room in the work environment. They occurred

during the same week, and sessions lasted from 40 to 60 minutes. We asked the participants to not shared details of the study to others outside of their workshop sessions.

At the beginning of each session, the first author acting as a moderator informed the participants about the aim of the study and the structure of the workshop. They were also asked for their consent, prior to gathering any data. Further, the moderator presented the Framework, followed by a few minutes for any questions/doubts to be asked by the participants.

Participants were instructed to bring their laptops, as we provided them with an electronic spreadsheet[2], which gathered data on their assessment, and also some demographics data.

With the moderator being present, each participant filled the spreadsheet individually. They were allowed to ask any further clarification and even to express their overall ideas to the colleagues in the same workshop session. We took notes of the participants' interaction with the moderator and among themselves. Next, the moderator explained the structure of the instrument and provided any further clarifications, as per the participants' needs.

Participants filled out the electronic spreadsheet individually. They were also allowed to ask the moderator for clarifications and to express their overall understanding of a given criterion to the colleagues in the same workshop session. We took notes of the participants' interaction with the moderator and between themselves.

3.1.5 Data analysis

To facilitate answering our research questions, we aggregated participants' answers into a shared data frame[3]. We further normalized the values by weighting the points assigned to each subcriterion taking into account the points allocated to its related main area. The equation for normalized importance is: $(a * n(a) * c^a)/1000$, where:

a is the amount of points assigned to a main area,
$n(a)$ is the number of subcriteria related to main area a, and
c^a is the amount of points associated to a particular subcriterion

As an example, given that a participant assigned 225 points to main area A, which contains six subcriteria; and 150 points to the subcriteria A1, the resulting normalized value for A1 is *(225 * 6 * 150) / 1000 = 202.5*.

The normalization standardizes the range of values over all the subcriteria, while at the same time considering the relative importance of the related main area. It also ensures that the probability of assigning values by chance is the same for all subcriteria. A similar procedure is employed by [22] for hierarchical voting analysis.

After computation, a given participant's relative importance values always adds up to 1000 points. This allows for the comparison between the relative importance assigned by different participants or participants' subgroups (e.g. PhD students).

[2]available at: `https://bit.ly/2InnyhG`
[3]available at: `https://bit.ly/2InmXMY`

Further, we provide an overview of the values assigned to each criterion from all participants through boxplots. This method for graphical representation of distributions gives a good indication of the dispersion of the data. Less dispersed distributions of importance values represent a higher degree of alignment among participants.

To interpret the results of the boxplots, we computed an importance threshold of 250, i.e. sum of all points (t = 1000) divided by the total number of main criteria (n = 4). Similarly to Rovegård, Angelis & Wohlin [21], we drew four distinct importance categories based on the characteristics of the distribution in relation to the threshold:

- **Unimportant (U):** the third-quartile is lower than the importance threshold
- **Lower importance (LI):** importance threshold is between the median value and the third-quartile
- **Important (I):** importance threshold is between the first-quartile and the median value; and
- **High importance (HI):** the first-quartile is higher than the importance threshold.

Finally, we calculated a priority list by ranking the quality dimensions according to the median value of the importance distributions. Such priority list represents the quality aspects most important, based upon participants' assessments.

3.1.6 Validity threats

Next we discuss a series of potential validity threats, using as basis the four aspects of the validity for case study research described by Runeson and Höst [14]. We also detail the procedures taken to handle the identified threats.

Construct validity: One potential threat relates to results bias if the data collection instrument is not designed carefully. To mitigate such threat, we opted to use the hierarchical voting prioritization approach, which has been successfully applied in several studies (e.g. [21,22]). These studies' reported experiences contributed towards the creation of our questionnaire. Prior to using the instrument, we tested it in a pilot with the participation of six subjects presenting similar research profile to our participants.

The quality dimensions are grounded on the work of Mårtensson et al. [4], which has been validated in a multidisciplinary context. To mitigate any misunderstandings relating to the framework, we presented and discussed the Framework and its elements in the beginning of the workshops.

Internal validity: Our case study aims to identify the participants' view of the quality dimensions for overall SE research. However, there is a limitation associated with the research preferences of each participant. For example, a researcher that conducts mainly experiment studies is likely to interpret the quality framework with a more positivist view. To mitigate the threat related to a narrow spectrum of SE research, we instructed the participants to consider the dimensions on a broader ESE context.

External validity: Our study was conducted in the particular context of the research practitioners of SERL Sweden. The candidates were selected based on

convenience and produced a quite small sample. From the 17 participants, 9 are PhD students, three Postdocs, three Assistant professors and only one Professor. As there are more junior than senior researchers, we could not assume that our participants represent the overall population of researchers in software engineering.

Our sample is reasonably heterogeneous with regard to their familiarity to different research methodologies and topics of investigation (see Section 3.2.1). Most of them has also experience in collaboration with industry, but only one is familiar to pure theoretical research. Such characteristics of the sample limit our discussion to the context of the observations.

Reliability: There is a potential threat related to the misinterpretation of collected data. To mitigate such threat, both data analysis and interpretation procedures were designed in advance, based on the works of Kuzniarz & Angelis [22], and Petersen, Khurum & Angelis [23]. Mainly, we normalized the importance values, prior to interpretation, given the 23 dimensions were not equally distributed across the four main areas. All the transformations applied over the data are provided in our data set2 so that other researchers can validate and replicate our results.

3.2 Results

3.2.1 Participants' Demographics

In total, 17 researchers participated in our case study, from which 9 are PhD students, three Postdocs, three Assistant professors and only one Professor. In average, they have been performing (and also reviewing and assessing) research in SE for more than 7.5 years, with a maximum of 30 years and a minimum of least than a year.

For the research type, most of the participants conduct empirical research in collaboration between academia and industry (9 out of 17 participants) or mostly in industry setting (3 participants). Just one researcher mentioned being familiar with pure theoretical research. Case study is the most prolific research method employed (by 12 participants), followed by systematic literature review (9), systematic mapping (6), interviews (6), survey-based research (5) and quasi-experiments (5). Other research methods the participants are familiar with include action research, archival research, design science, experiment, focus group, meta-analysis, observations, simulation, and thematic analysis.

Their research covers a wide range of SE topics, such as Software Testing (7 participants), SE Management (3), SE Process (3), Requirements Engineering (3), Software Construction (3) among others. The participants have published an average of 7.5 journals and 13.8 conference papers. The preferred venues for publication of their work are IST - Information and Software Technology (13 participants), JSS - Journal of Systems and Software (11), and EMSE - Empirical Software Engineering (6). The preferred conferences are ESEM - Empirical Software Engineering and Measurement (6 participants) and EASE - Evaluation and Assessment in Software Engineering (4).

3.2.2 RQ1a) Importance of the Quality Dimensions

Boxplots for all quality dimensions are shown in Figure 2 (four main areas) and Figure 3 (main areas' subcriteria). The colours used in all the boxplots relate to the four main areas, as follows: red for credible (A), blue for contributory (B), green for communicable (C) and yellow for conforming (D).

Fig. 2 Distribution of importance values assigned to the four main areas.

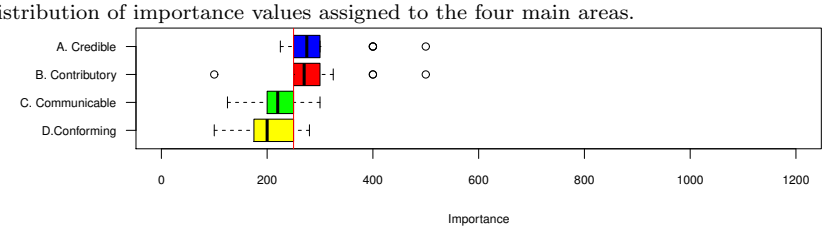
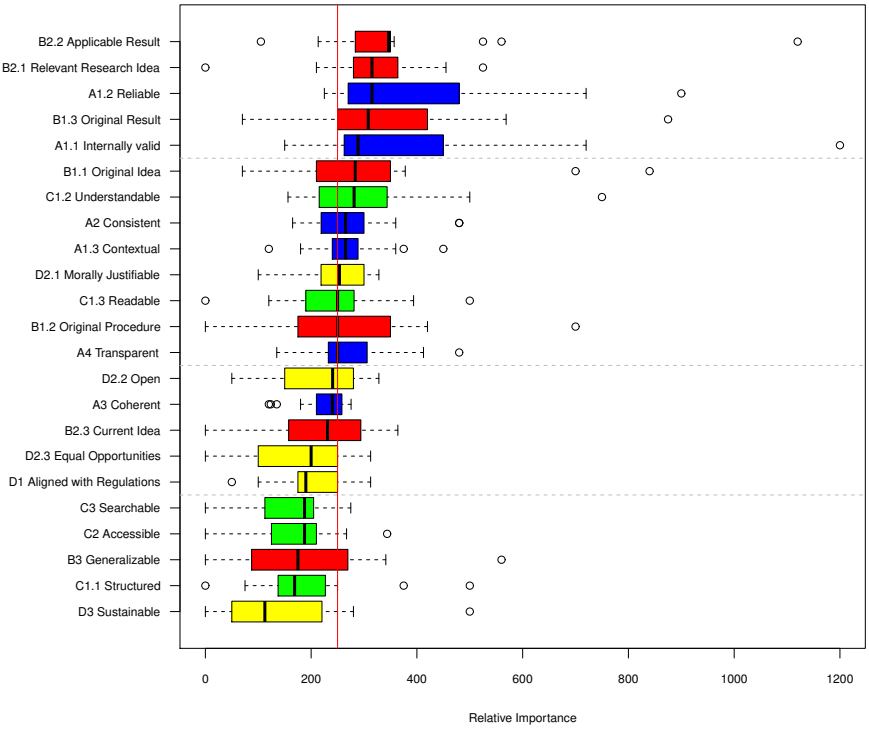


Fig. 3 Distribution of relative importance assigned to the quality subcriteria. Relative importance values are normalized by weighting the points assigned to each subcriterion by the points allocated to its related main area.



Boxplots in Figure 2 show the distribution of importance assigned to each main area, whereas Figure 3 show the distribution of relative importance (weighted according to the related main dimension) assigned to each subcriteria. Boxplots are displayed according to the priority list, i.e., higher medians are shown on the

top, and lower medians at the bottom. Further, a threshold line is drawn vertically over the plot, thus allowing us to divide it according to importance categories, from High (HI) at the top, to Low (U), at the bottom.

Figure 2 shows that the main areas credible (A) and contributory (B) are ranked as more important than the other two main areas, with a median above the threshold; followed by communicable (C), and finally conforming (D). With regard to the main areas' subcriteria, except for A3, all subcriteria for the main areas A have medians above the threshold level, thus clearly suggesting a higher level of importance. Most of the subcriteria for the main area B also have medians above the threshold level, when compared to the subcriteria for the other two main areas (C) and (D). The subcriteria C1.2, D2.1 and C1.3 are the only three subcriteria from the main areas (C) and (D) that have medians at or above the threshold, respectively.

Five dimensions were considered of higher importance to the participants: B2.2 applicable results, B2.1 relevant research idea, A1.2 reliable, B1.3 original result, and A1.1 internally valid, respectively. It is important to notice that this set comprises 2 out of 3 dimensions related to the rigorous and relevant subareas.

Lower importance dimensions are mostly associated with communicable and conforming areas. In particular, the unimportant (U) subcriteria are related to these two areas. The only subcriteria associated to the communicable area rated important is C1.2 understandable and C1.3 readable. From the communicable area, D1.1 morally justifiable is the only subcriterion rated important.

3.2.3 RQ1b) Participants' Alignment

The characteristics of the distributions of importance in Figure 2 (four main areas) and Figure 3 (main areas' subcriteria) are also indicators of the level of alignment between participants. Shorter boxes and tail length denote a higher level of agreement over the number of points given, whereas larger boxes would be the opposite.

Besides the visual representation, characteristics of the distribution are useful to identify levels of alignment. In particular, measures of variation, such as the standard deviation, quantifies the dispersion of a distribution. A low standard deviation means that the assigned values of importance are concentrated close to an expected value, thus denoting alignment. Table 1 summarizes the characteristics of the distributions that could be used to determine alignment.

A close alignment is represented in the boxplot by smaller boxes and short whiskers, e.g. A1.3 contextual (σ = 37.36) and A3 coherence (σ = 44.48). They are consistently ranked (important and lower importance, respectively) across the quality dimensions. This implies that the views towards the importance of these criteria are more aligned and do not change much within the participant group.

As examples of misalignment, the subcriteria A1.1 and A1.2 (σ = 254.09 and 192.51, respectively) are widely distributed, thus suggesting that the participants do not share a common perception about the importance of these dimensions. They are also skewed to the right, indicating that the assigned importance tends to the lowest values. Similar misalignments exist with relation to several contributory subcriteria, e.g., B2.2 applicable result (σ = 218.03) and B1.1 original idea (σ = 190.46). The distribution of importance ratings for these subcriteria is highly influenced by the outliers.

Table 1 Summary of the characteristics of the distributions presented in Figure 3. Each subcriterion is presented in a row, followed by its importance category; minimum, median and maximum values for the distribution; and the standard deviation (σ).

Sub-criterion	Category	Min	Median	Max	Std. Deviation (σ)
A1.1 Internally valid	HI	150	288.8	1200	254.09
A1.2 Reliable	HI	225	315	900	192.51
A1.3 Contextual	I	120	265.2	450	76.73
A2 Consistent	I	165.00	265.2	480	91.07
A3 Coherent	LI	120	240	275.6	51.55
A4 Transparent	I	135.00	250	480	82.33
B1.1 Original Idea	I	0	283.5	840	190.46
B1.2 Original Procedure	I	0	250	700	160.21
B1.3 Original Result	HI	70	308	875	187.06
B2.1 Relevant Research Idea	HI	0	315	525	114.63
B2.2 Applicable Result	HI	105	346.5	1120	218.03
B2.3 Current Idea	LI	0	231	364	106.58
B3 Generalizable	LI	0	175	560	148.47
C1.1 Structured	U	0	168.8	500	115.40
C1.2 Understandable	I	156.20	281.2	750	144.24
C1.3 Readable	I	0	250	500	112.13
C2 Accessible	LI	0	187.5	343.8	89.83
C3 Searchable	LI	0	187.5	275	76.42
D1 Aligned with Regulations	LI	50	190	312.5	74.49
D2.1 Morally Justifiable	I	100	253.5	328.1	68.98
D2.2 Open	LI	50.00	240.6	328.1	91.80
D2.3 Equal Opportunities	LI	0	200	312.5	89.97
D3 Sustainable	U	0	112.5	500	126.79

Outliers are represented by small circles beyond the bounds of the whiskers. Herein they point our importance rates assigned by a participant that does not agree with most. The outliers are very common in our boxplot of the normalized importance values (illustrated in Figure 3), except for four dimensions (D2.1, D2.2, B2.3, and C3, respectively). It is interesting to note that outliers are more common towards higher values of importance.

Several distributions have a lower limit equals to zero (0). Explicitly, several subcriteria on the lower importance (LI) and unimportant (U) categories fit this description. This means that the criterion is considered non-important (assigned value = 0) by at least one participant. In particular, B3 generalizable, C2 accessible and D3 sustainable are assigned a non-important for more than one participant. About B3, one of the participants commented: *"[it] depends on the type of the study, e.g., case study research is not generalizable at all outside the context."*

4 Focus Group

4.1 Method

We employed a focus groups methodology to gather experts' opinion regarding the importance of research quality dimensions, where such dimensions are also based upon the model of research quality by Mårtensson et al. [4]. Further, we summarize

our research method according to the guidelines provided by Kontio, Lehtola & Bragge [24].

4.1.1 Defining the research problem

Our second study aimed to bring together the perceptions of senior ESE researchers towards the dimensions of research quality as per the abovementioned framework. Besides the relative importance of the dimensions, we also wanted to promote discussions and knowledge building on what constitutes "research quality". The following research questions geared this study:

RQ2a What are the most important research quality dimensions in SE research, as per the participants?

RQ2b What is the level of alignment between the different participants with regard to their assessment of the importance of quality dimensions for SE research?

RQ2c To what extent the model of research quality proposed by Mårtensson et al. [4] is relevant for software engineering research?

4.1.2 Selecting the participants

The study was conducted within the context of the International Software Engineering Research Network (ISERN), which is a community of SE researchers (in particular senior/experts) and practitioners who apply an evidence-based paradigm to SE research. These SE researchers also have knowledge on methodological research, and commonly employ (and some also propose) different research methods to conduct empirical studies in SE.

Further, this community also contributes with steering the research community towards the quality of research, and discussions on the subject have led to publications that raise awareness of the importance of the quality and appropriateness of the methods used in the SE empirical studies, e.g. [25,26,27,28].

ISERN holds annual meetings as part of the Empirical Software Engineering International Week (ESEIW). Thus, we proposed the ISERN organization to conduct a focus group study during their meeting. Further, we sent email invitations to ISERN participants and also advertised the study during the ESEIW.

4.1.3 Planning and conducting the focus group session

We designed the focus group to fit a session lasting 1.5 hours, consisting of six parts[4]. The first author acted as mediator, fostering discussion among the participants, while the second author served as an observer, taking notes and recording additional data. The participants were allowed to ask questions to the moderator and to the other co-participants during the whole session.

We conducted two sessions, the first one with eight participants and the second one with four participants. Each session started with an introduction in which we provided an overview of our objective, i.e. to discuss what constitutes "research quality" according to their perspectives. We also provided practical instructions

[4]Script for focus group sessions is available at: `https://bit.ly/2REKUmC`

for the focus group session and asked participants for verbal consent for audio recording.

Further, we presented the conceptual model of research quality [4], describing its components and structure. This model was used as a starting point for the discussion, and we carried out a prioritization exercise similar to the one done in our case study (see Section 3.1).

Next, we provided participants with a whiteboard in which they could share their main opinions visually. We asked them to use green and red post-its to point out the most and the least important subcriteria for each of the four main areas. This representation was used to guide the discussions in the last part of the session.

Finally, we moderated a semi-structured discussion covering all the dimensions of the conceptual model. In particular, we focused the discussions on potential misalignments, i.e. whether dimensions were assigned high (green post-its) and low (red post-its) importance by different participants.

4.1.4 Analysis

The focus groups sessions were documented using four different media: 1) filled out prioritization questionnaires, 2) whiteboard containing the visual representation of the highest and lowest important dimensions, 3) audio recordings of the sessions and their transcripts, and 4) notes from the observer.

Out of 12 participants, 9 filled out the prioritization questionnaires. To analyse the importance of the quality dimensions, we employed a hierarchical voting analysis similar to the one used in our case study (see Section 3.1). The prioritization values of all participants were aggregated into a shared data frame[5]. Later, we computed the relative importance values, and the subcriteria with the values assigned to the four main dimensions.

To interpret the results, we used the same importance threshold and categories described in Section 3.1. The normalized values of importance and distributions are therefore comparable to the results of our case study. We also used the visual representation of the post-it board to triangulate the prioritization results.

The transcripts of the focus group discussions were later analysed by the first and second authors, using pattern-matching and vote counting [24]. The opinions of the participants were summarized into a short sentence and then mapped to related dimensions. Further, we counted the participants who mentioned or agreed with such an idea. We also identified cases in which conflicting ideas were provided, e.g. whether to merge A4 Transparent to another dimension or to keep it separate.

The analysis resulted in the identification and categorization of opinions with regard to the dimensions of quality. It also provided insights about improvement needs of the conceptual model with regard to dimensions that are redundant, dependent or poorly described. Such suggestions were also organized and used to revise the model.

We also used the observer's notes to cross-check participants' opinions, as per the moderator's recollection. Moreover, the post-it board allowed us to confirm whether the comments from participants were related to a divergent opinion regarding a quality dimension. This further triangulation was used to support the decision making regarding the revision of the conceptual model.

[5]available at: `https://bit.ly/2BGHpTr`

4.1.5 Validity Threats

The potential threats to the validity of our focus group results are discussed with regard to four categories, as suggested by Runeson and Höst [14]:

Construct validity: The Framework of research quality has been created and evolved during workshops with experts from multiple research fields [4]. To minimize any potential threat related to the incorrect interpretation of the quality dimensions, we presented and discussed the Framework's structure and its components during the focus group session.

Further, the participants could also discuss between themselves, so aiming to have a common understanding of the conceptual model's dimensions and subcriteria. The moderator also suggested participants share conflicting opinions, so seeking a complete interpretation of the dimensions through the divergent thinking process.

Internal validity: To ensure the trustworthiness of our study, the first author designed a focus group protocol, later evolved by discussing it with the second author. Finally, the third and fourth authors reviewed the design by reflecting on the understandability of the questions and the accuracy of data collection. To reduce researcher bias, during the focus group session, the observer also acted as moderator's referee, keeping a record of any action that could threaten the nature of the discussion. None of such actions were identified.

In general it is difficult to keep balanced group dynamics in focus groups sessions. We employed a semi-structured design to ensure that all the participants had the opportunity to express their opinion. Later on, we assessed the number of suggestions provided by each participant, to identify the ones that influenced the activity more. Participants were self-selected from the ISERN members attending their annual meeting.

External validity: Our focus group gathered opinions from a small group of experts who are senior researchers and have the expertise to judge the topic of the discussion, i.e. quality of research in software engineering. They are also members of a leading community in software engineering research.

To become an ISERN member, one has to apply (with their institution) and evaluated by a senior committee who decided whether their application will succeed or not. Although our sample does not represent the diversity of the SE research population, it supports the transferability of our findings by expertise, i.e. our results are drawn from the opinions of experts who are likely to influence the broader community.

Reliability: The discussions and interaction of the participants provided us with rich qualitative data reflecting their opinions. To ensure that the opinions were correctly understood by the researchers, the moderator often paraphrased the participants to obtain confirmation. Later, the analysis and interpretation of the qualitative data were carried out according to the study design. The moderator summarized and categorized the opinions, further reviewed by the observer.

4.2 Results

4.2.1 Participants' Demographics

Twelve members from the ISERN participated in our focus group study. Most of them are academics: professors (5 participants), associate professor, assistant professor, and post-doctoral researcher (1 participant each). Two participants work in other research positions, i.e. research scientist and head of research department. Finally, one of the participants works on the industry as IT consultant.

The participants have published an average of 17.4 journals and 63.5 conference papers, well above the participants in the case study. Just five of them provided us detailed information regarding their research preferences and previous experiences. Research methods mentioned by these participants include case study research, design science, interviews, systematic literature review, systematic mapping, experiments, quasi-experiment, statistical and qualitative analysis.

4.2.2 RQ2a) Importance of the Quality Dimensions

Here boxplots were also used to display the distribution of values (level of importance) assigned to each quality dimension (see Figures 4 and 5). We ordered the subcriteria by the median values and used the threshold (vertical red line) to classify them according to importance. The boxes' colours represent the main areas to which the subcriteria are grouped, i.e. red for credible (A), blue for contributory (B), green for communicable (C) and yellow for conforming (D).

One can notice some similarities with our case study's results, such as the overall importance assigned to the main area A, followed by B, and then C and D. Despite that, the data shows that the ISERN experts favour the subcriteria related to credible main area over all the others, except for A3 Coherent, ranked as lower importance. Two subcriteria (A1.1 and A4, respectively) are considered of high importance (HI) according to our categorization, as their boxes are placed above the threshold.

With regarding the contributory aspect, three subcriteria (B3, B2.2, and B2.1) have a median above the threshold, thus categorized as important (I). The four other subcriteria (i.e., B1.3, B2.3, B1.1, and B1.2) have medians below the threshold. Specifically, the third quartile of B2.3 current idea below the threshold, thus unimportant (U) according to our categorization. All the subcriteria related to communicable (C) and conforming (D) areas have medians below the threshold. It worth mentioning that the values are obtained by HCV. Thus the values assigned to the main areas impacts the final result for each related subcriteria.

These boxplots aggregate and summarize all participants' assessments, without being specific about any individual assessment. Exceptions to this are the outliers, i.e. data points that are abnormally far from all other participants. The outliers are less frequent than our case study, implying a better alignment overall. It is important to note that the sample size of the focus group is smaller (only 10 participants) and thus likely to impact on the distribution of quantitative data.

The extent of the bars and the whiskers does point conflicting opinions among the participants. This is the case for B2.1 relevant research idea, that is ranked as the most important at all for some participants and as low important by others.

Fig. 4 Distribution of importance values assigned to the four main areas.

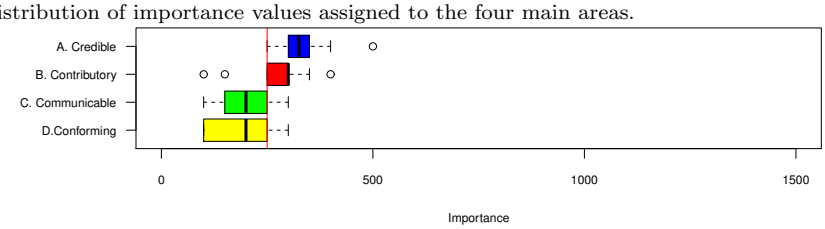
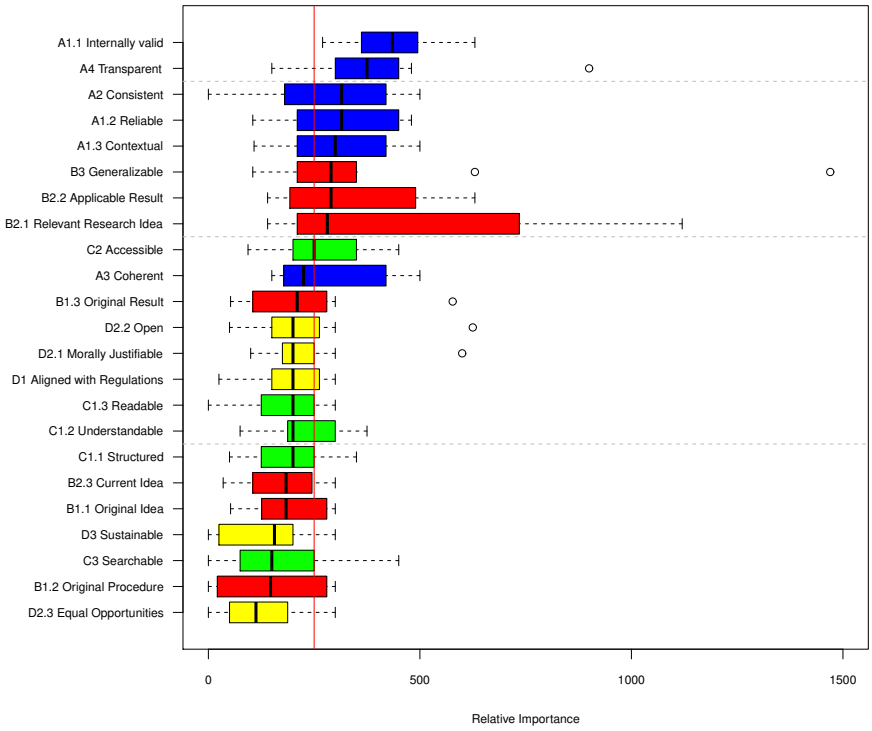


Fig. 5 Distribution of relative importance assigned to the quality subcriteria. Relative importance values are normalized by weighting the points assigned to each subcriterion by the points allocated to its related main area.



This potential misalignment was the focus of our discussions during the focus group and provided us deeper insights regarding the participants' opinions.

4.2.3 RQ2b) Participants' Alignment

Following the prioritization exercise, we carried out another activity aimed at assessing the same four main research quality areas and their corresponding subcriteria; however using another technique. We displayed on a whiteboard the four main areas and corresponding sub-criteria, and ask participants to pick those areas/sub-criterion that they saw as most important and lest important. To differentiate between these two types of decisions, participants were given green and

red post-its – green symbolising most important and red least important. Figure 6 displays the results for both groups – Group 1 (8 participants), and group 2 (4 participants) are shown on the left and right hand sides, respectively.

Fig. 6 Representation of a whiteboard describing the most and least important dimensions, according to the focus group sessions. Each green dot represents a sub-criterion assessed as highly important by a participant; conversely, a red dot represents a sub-criterion assessed as least important.

Note that there are more dots in each area than the number of participants; this occurred because some participants considered more than one sub-criterion as equally important. Also note that this is a different assessment in which there was no normalisation of results; therefore, the assessment shown in the whiteboard does not map entirely to the normalized values of importance provided in Figure 5. In other words, the whiteboard presents the view of the participants with regard to a particular area (e.g., credible) and does not account for the importance of each sub-criteria weighted according to its main area, as described in Section 3.1.3.

Table 2 Summary of the answers from both focus group sessions at the whiteboard. In the columns labelled Green and Red we account for the amount of markers in relation to the group size, i.e. number of participants. The rigghtmost column describe our interpretation of the findings, according to three categories.

	Focus Group 1		**Focus Group 2**		
Sub-criteria	Green	Red	Green	Red	Interpretation
A1.1	100% (8/8)	-	75% (3/4)	-	ABG
A1.2	87.5% (7/8)	-	25% (1/4)	50% (2/4)	AWG(1), MWG(2)
A1.3	50% (4/8)	-	50% (2/4)	50% (2/4)	MWG(2)
A2	12.5% (1/8)	-	50% (2/4)	25% (1/4)	MWG(2)
A3	12.5% (1/8)	-	-	50% (2/4)	
A4	37.5% (3/8)	12.5% (1/8)	25% (1/4)	-	
B1.1	25% (2/8)	50% (4/8)	-	25% (1/4)	MWG(1)
B1.2	25% (2/8)	62.5% (5/8)	-	50% (2/4)	MWG(1)
B1.3	37.5% (3/8)	50% (4/8)	-	25% (1/4)	MWG(1)
B2.1	75% (6/8)	-	75% (3/4)	-	ABG
B2.2	50% (4/8)	-	-	-	
B2.3	50% (4/8)	-	-	25% (1/4)	MBG
B3	12.5% (1/8)	-	25% (1/4)	25% (1/4)	MWG(2)
C1.1	62.5% (5/8)	25% (2/8)	25% (1/4)	25% (1/4)	ABG, MWG(both)
C1.2	50% (4/8)	25% (2/8)	50% (2/4)	-	MWG(1)
C1.3	50% (4/8)	25% (2/8)	100% (4/4)	-	AWG(2)
C2	12.5% (1/8)	-	50% (2/4)	-	
C3	12.5% (1/8)	-	25% (1/4)	50% (2/4)	MWG(2)
D1	25% (2/8)	25% (2/8)	50% (2/4)	25% (1/4)	ABG, MWG(both)
D2.1	75% (6/8)	12.5% (1/8)	50% (2/4)	-	MWG(1)
D2.2	50% (4/8)	12.5% (1/8)	-	-	MWG(1)
D2.3	37.5% (3/8)	37.5% (3/8)	-	50% (2/4)	MWG(1)
D3	25% (2/8)	25% (2/8)	-	50% (2/4)	MWG(1)

An excellent example of this difference is related to the sub-criteria B2.1 relevant research idea. It received a large number of green markers (see Figure 6), but it is not ranked as the most important sub-criteria of the contributory dimension (Figure 5). In particular, the wider distribution of its box and right whisker suggests misalignment on the opinion of participants with regarding B2.1.

In order to summarize the results of the whiteboard, we applied vote counting to the markers, dividing the total of reds and green markers by the total number of participants so to obtain a normalized view, as detailed in Table 4.2.3. The interpretation of the results of the whiteboard exercise were categorized according to three different patterns, as follows:

ABG) Alignment between-groups: represents whether similar views (either in agreement or conflicting) emerged in both groups.

AWG) Alignment within-groups: represents whether most of the group participants agree on the importance of a particular sub-criteria.

MBG) Misalignment between-groups: represents whether there are conflicting views of importance between participants from different groups.

MWG) Misalignment within-group: represents whether conflicting views of importance exist within participants of a group.

With regarding the alignment between-groups (ABG), we identified two major cases. First, A1.1 internally valid is considered important by most of the participants in both groups (more than 90% of participants). Second, B2.1 relevant

research also received a lot of green markers from both groups. This result suggests a preference for methodological strictness and practical relevance.

In some other cases, we noted aligned views of importance within a group (AWB), but that is not shared by the participants of the other groups. Most of the participants in group 1 considered the subcriteria A1.2 reliable as important. Participants from group 2 are more aligned towards C1.3 readable, the better ranked sub-criteria related to main area C communicable.

Misalignment between-groups are not easily explained, as the participants from different groups did not discuss the issue between them. We identified a relevant MBG case related to B2.3 current idea, within the contributory area.

Misalignments within-group (MWG) were more common, and they were also the ground for discussions during the focus group sessions. With regarding focus group 1, most of the MWG were identified concerning the original subcriteria (i.e., B1.1 to B1.3) and the main area conforming (i.e., D1 to D3). In relation to group 2, the MWG occurred about three subcriteria of the credible main area (i.e., A1.2, A1.3, and A2), two of the conforming dimension (C1.3 and C3) and another one concerning B3. Finally, two MWG were identified in both of the groups (C1.1 and D1), thus denoting alignment between-groups.

4.2.4 RQ2c) Summary of the participants' opinions

Misalignments within-group were also ground for the discussions during the focus groups sessions. Whenever there was an MWG with regard to a main area's or subcriterion's importance (i.e., both green and red markers were present), participants were invited for discussion. The goal of such discussion was not to reach consensus but to understand in more detail their rationale. Notes were taken during such discussions; we also recorded the focus group sessions and used it to complement the notes[6].

Besides the relevance of the dimensions, participants discussed how such model should be operationalized, and what should be the intent behind using such model to assess research quality. The group 1 discussions focused upon the context to which such quality dimensions apply. In particular, the opposing views of academia vs. industry, quantitative vs. qualitative research, research practice vs. research report, and the different philosophical stances (e.g., positivism, pragmatism) were mentioned.

Both groups also pointed out descriptions of several dimensions that, in their view, were not clear. This was also true in relation to the glossary (see Appendix A.2), which is provided in the original paper, and is supposed to facilitate understanding. Some of the participants suggested improvements to the phrasing and also to the model's structure.

Both groups criticized the model's structure. Further, one of group 1's participants also mentioned that the process to cluster or aggregate the dimensions was questionable. Participants also pointed out redundancies (e.g., between C1.2 understandable, and C1.3 readable; dependencies (e.g., A1.2 reliable depends on A4 transparent), and overlaps (e.g., B3 generalizable seems to belong both to contributory and credible areas).

[6] A summary of the suggestions is compiled in `https://bit.ly/2CiXqOQ`

Finally, both groups also believed that the relationships between some dimensions were not explicit, and the dependencies should be made clearer. One participant suggested the use of pass-only standard, shaping the model into a series of decision points. We organized such suggestions and proposed a revised model of research quality (see Section 5.3).

5 Discussion

5.1 A Method to Assess the Importance of Research Quality

Our research approach used the method to determine the alignment of stakeholders proposed by Barney and Wohlin [12]. The method, originally designed for software quality standards, performed well under the circumstances of our study. In particular, it matches Mårtensson et al. [4] recommendations, in which the dimensions of quality should be discussed, weighted and prioritized by experts on the field it is intended to be applied.

Researchers willing to apply a similar approach to assess the alignment of different groups of experts are encouraged to follow our methodology. Therefore, we describe here the process we employed:

Select a case context. First, it is essential to ensure that the context in which the process will be applied is sufficiently focused. In this article, we described two distinct applications, the first with researchers from SERL Sweden, and the second within the ISERN annual meeting (see Sections 3.1 and 4.1 for details). In both cases, we investigated the generic standards of research quality in ESE, but more specific topics (e.g., related to a particular philosophical stance or activity in the research process) are also interesting of an investigation.

Identify key stakeholders. To ensure that our candidates are interested in the investigated topic, we advertise the research idea and provided means for self-selection. The participants on both cases covered a wide range of research competencies and provided different perspectives regarding the topic.

Select/tailor a quality model. We selected the multidisciplinary framework for classification of the quality of research proposed by Mårtensson et al. [4] to represent the aspects of research quality. Alternative models, describing different quality concepts are available, e.g. [8,9,10].

Develop a data collection instrument. We developed the data collection instrument as an excel spreadsheet[2] grounded on previous works [22,23]. The subjects assigned importance values using hierarchical cumulative voting (HCV), and the assigned value for each subcriterion was weighted according to the values assigned to its related main area.

Conduct data collection. We conducted workshop sessions to collect the selected participants' opinion. In the case study, the workshops took place in a meeting room at the work environment and last from 40 to 60 minutes in length. In the focus group, we conducted two sessions in the ISERN annual meeting, lasting 90 minutes each. The focus group sessions also included a whiteboard representation of the participants' reasoning and round-robin discussions.

Analyze the results. A detailed explanation of the data analysis procedure is given in [12]. We aggregated the relative values of importance into a shared data frame and used descriptive statistics to present the results. We interpreted the

data according to predefined categories (see Section 3.1) and identified relevant patterns regarding alignment and misalignment.

Present the results to participants. The results should be presented to participants asking them for responses to follow-up questions. After the case study's workshops, we showed the preliminary results for the participant's audience and collected their feedback to improve our study. The feedback was incorporated into the focus group design, in which a revised model of research quality for software engineering was produced (see Section 5.3). Finally, we sent the resulting model to the participants and asked them for feedback.

5.2 Alignment with Regard to the Importance Values

In this article, we presented two distinct studies (i.e. a case study and a focus group), each of which aiming to collect the views of importance from different stakeholder groups. This resulted in two distributions of relative importance (see Figures 3 and 5). Although similarities can be noted, they represent the particular views of two different groups.

During the case study, we determined the alignment according to the distributions of assigned importance values, i.e. the length of boxplots and whiskers, and the relative position of its central point with regarding a threshold. We identified interesting patterns, but could not explain the reason for such. Based on the feedback provided by participants, we designed a more comprehensive study.

As a next step, we employed a focus group approach to further investigate the perspective of experts regarding the conceptual dimensions of quality. The importance values were summarized by the participants themselves, providing a visual description of potential alignments and misalignments. Further discussions elicited different views regarding the importance of quality dimensions, the conceptual model and its applicability.

The mixed media we adopted in the focus groups provided richer evidence in comparison to the prioritization questionnaire alone. Additional qualitative data (i.e. group session transcripts and notes) provided a deeper understanding of the alignment with regard to the importance of quality dimensions.

Participants also pointed out dependencies among the dimensions. We would like to conduct further studies to explore the interdependence between quality dimensions and identify potential trade-offs under different perspectives. We also encourage researchers willing to apply and evolve our approach to report the results and experiences from future applications.

5.3 Revised Model of Research Quality

During the focus group session, we explicitly asked the participants if they think the model proposed by Mårtensson et al. [4] express the dimensions of research quality. Most of them agreed, but they mentioned a few concerns regarding:

1. The structure of the model;
2. The description of the dimensions; and
3. Applicability of the model for research appraisal.

We further discussed these issues and elicited suggestions from the participants on how to address such challenges. To solve issues 1 and 2, we revised the model based on the suggestions and propose new descriptions grounded it the literature.

Our update mainly reflects the suggestions of the focus group session. We summarized the comments and addressed one dimension at a time, assessing whether the suggestions are essential. Dimensions that are classified as highly important or unimportant (Section 4.2.2) and the ones presenting misaligned opinions (Section 4.2.3) were given extra attention.

In some cases, suggestions from different participants conflicted, so we applied a decision-making process based on vote counting. We used the transcripts to assess how many participants mentioned and agreed with a particular suggestion. We kept the four main areas, as none participant understood that this is an essential issue to the model. Figure 7 shows the revised structure of the model, which include the following updates:

- *A1.2 Reliable:* broken down in two: A1.2a Methodological rigorous and A1.2b Reproducible. Methodological rigorous is described as "The research method is correctly used for its intended purpose" [9, 29, 30]. Reproducibe is "The findings are likely the same if the research is repeated, e.g. by independent researchers, at a different time, in a different place" [31].
- *A5 Coherent:* merged into A4 Consistent; the new merged dimension is defined as "Consistent: The research problem, research questions and research methodology are linked" [30].
- *B1.1 to B1.3:* merged into B1 Original. The new Original dimension is described "The research addresses a new problem by means of original theory, methods and context" [32].
- *C1.2 Understandable:* merged into C1.3 Readable, which is now defined "The research documentation is able to be understood (in terms of tone, style, structure and semantics) bySE professionals" [9]; and
- *C5 Searchable:* merged into C4 Accessible; which is now "The research report is available and is easily found by the potential audience, in particular outside of academia" [10].

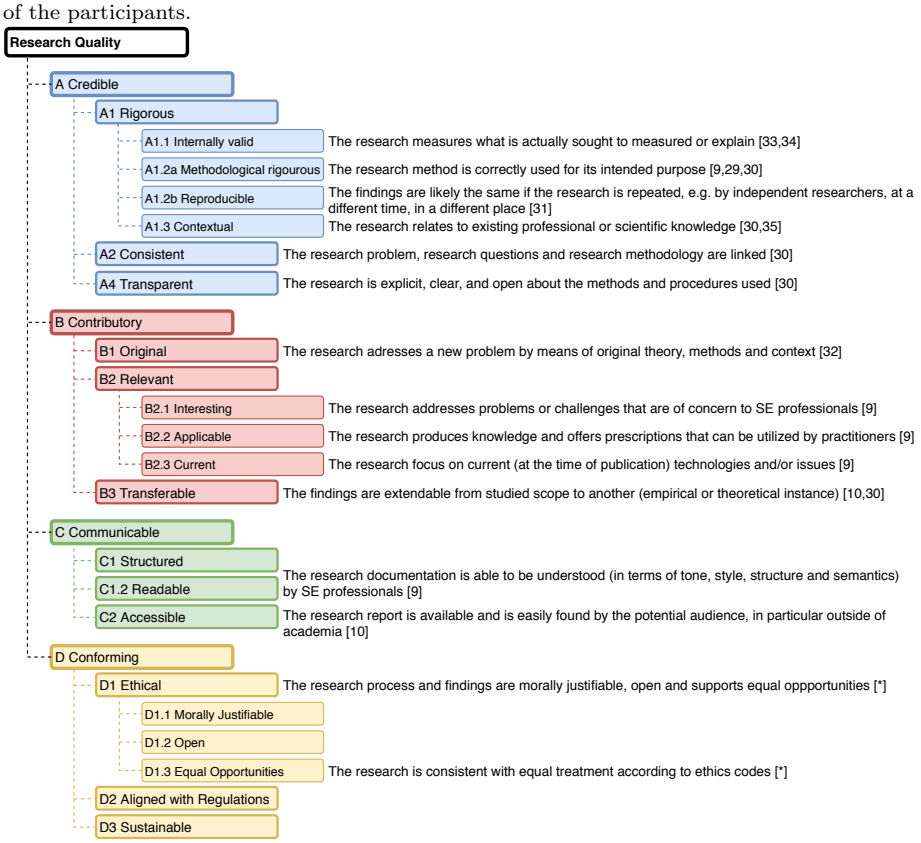
We also incorporated new descriptions for the dimensions that were mentioned vague or hard to understand. Fourteen out of the 26 items figuring the revised model have been rephrased as requested by the focus groups' participants, as shown in Figure 5. In addition to their suggestions, we look up to methodological papers describing such quality dimensions to compose a new description.

We systematically identified a set of nine papers supporting a new description of the requested items, six of which are cited by Mårtensson et al. [4]. The papers cover a series of domains, such as Software Engineering [9, 10, 33, 29], Research Methodology [30, 34], Information Systems [32], Social Sciences [31] and Spatial Planning [35]. Besides that, we adopted two suggestions of the participants.

5.4 Using the Model to Appraise Research Quality

The original conceptual model [4] is intended to foster the development of research practice in multiple academic fields, by promoting a shared understanding of the quality dimensions. Suggested applications include appraising research reports,

Fig. 7 Revised model of research quality grounded on the results of the focus group. The updates comprise a new hierarchical structure and novel descriptions of fourteen dimensions, listed on the right. Items marked with an asterisk (*) were rephrased according to suggestions of the participants.



assessing research grants applications and support the creation of standards for science and innovation.

Before operationalization, the authors suggest that the model should be validated in the domain it is intended to be applied [4]. By conducting such validation, this study resulted in a revised model of research quality. To demonstrate how the model can be applied to research practice, we design an exemplary scenario:

Assuming we are willing to employ a checklist instrument to appraise the quality of research, we would like to investigate the comprehensiveness of the checklist with regard to the quality dimensions.

We opted for using an unified checklist [36] that combine a series of recommendations for experiments and case study research. We employed a coding process using the dimensions in the revised model of research quality (see Figure 7) as a codebook. Table 1 summarizes the checklist items and the dimensions we assigned to each of them. A more extensive description of the checklist items is provided by Wieringa [36].

Table 3: Unified Checklist for Empirical Research in Software Engineering by Wieringa [36]. The rightmost column shows the categorization of the checklist questions according to the revised model of research quality.

ID	Question	Criteria
	Research problem investigation	
U1	What is the higher-level engineering cycle?	A1.3, B1
U2	Knowledge goal in that cycle?	A1.3, B1
U3	Conceptual model of the phenomena?	A1.3, A2
U4	Conceptual model validity? (including construct validity)	A1.1, A2
U5	Unit of study (population)?	A1.1
U6	Research questions?	B2.1
U7	Current knowledge?	B2.3
	Research design	
U8	Unit of data collection? (sample, model or case)	A1.1, A1.2a
U8.1	Acquisition?	‖
U8.2	Structure?	‖
U9	Treatment of unit of data collection?	A1.1, A1.2a
U9.1	Treatment specication?	‖
U9.2	Treatment assignment?	‖
U9.3	Treatment plan?	‖
U9.4	Treatment instruments?	‖
U10	Measurement of unit of data collection?	A1.1, A1.2a
U10.1	Measurement procedures?	‖
U10.2	Measurement instruments?	‖
U11	Kind of reasoning? (statistical or case-based)	A1.2a
	Research design validation	
U12	Validity of unit of data collection?	A1.1
U12.1	External validity?	B3
U12.2	Ethics?	D1
U13	Validity of treatment?	A1.1
U13.1	Instrument validity?	‖
U13.2	External validity?	B3
U13.3	Ethics?	D1
U14	Validity of measurement?	A1.1
U14.1	Validity of measurement procedures?	A1.2a
U14.2	Instrument validity?	A1.1
U15	Validity of reasoning?	A1.2a
U15.1	Conclusion validity?	‖
U15.2	Internal validity?	A1.1
	Research execution	
U16	Unit of data collection?	A4
U16.1	Acquisition?	‖
U16.2	Quality?	‖
U16.3	History?	‖
U17	Execution of treatment?	A1.2b, A4
U18	Execution of measurements?	A1.2b, A4
U19	Availability of data?	A4
U20	Provenance of data?	A4
	Results evaluation	
U21	Data?	A4, C1
U22	Observations?	C1
U23	Explanations?	C1
U24	Answers to research questions?	A1.1
U25	Generalizations?	B3
U26	Limitations?	A1.2a
U27	Contribution to knowledge goals?	B2.2

Continued on next page

Table 3 – *Continued from previous page*

ID	Question	Criteria
U28	Contribution to engineering goals?	B2.2

The checklist is structured in five sections mapping the phases of the empirical cycle [36]. The first section focus on the conceptualization of the study and the definition of the research problem to be investigated. Quality aspects herein are focused on rigorous method (A1), original and relevant contributions (B1 and B2).

The research design section concerns the proper use of research methods (A1.2a) to correctly measure the phenomenon (A1.1) it is sought to investigate. The next section addresses the validation of such design, and thus cover similar aspects, also, to ensure that the results are generalizable or transferable to other contexts (B3).

Research execution is focused on trustfully documenting the research process (A1.2a), including any unexpected event that could influence the results. Transparency (A4) is required for further assessment and replication of the study. Finally, the results evaluation section focus on communicating the research (C1) aiming to provide a meaningful contribution (A2.2) to the field.

The categorization of the checklist includes all four main areas of research quality, with different degrees of coverage. Areas (A) credible and (B) contributory are more expressed than (C) communicable and particularly, (D) conforming. In particular, questions addressing the rigorous aspect (A1) are plenty, characterizing the checklist towards the methodological quality of research.

About the communicable dimension, the checklist does not address the readability and accessibility of the report. With regard to the conforming aspect, only D2 ethics is explicitly mentioned, with no specification of its subcriteria. D1 alignment to regulations and D3 sustainability are not covered.

The categorization of the checklist items is consistent with the results of importance assessment in both our case study and focus group (Sections 3.2 and 4.2). In this sense, one can assume a match between the proposed checklist instrument [36] and the views of importance regarding the research quality.

We also identified gaps related to the communicable and conforming dimensions of research quality. On the one hand, these aspects are perceived as secondary, as they less concerned the research practice. On the other hand, their importance is grounded on the means researchers use to reach the potential audience and an ongoing debate on the ethical/moral orientation of research.

6 Conclusion

This paper reports a mixed method empirical study to validate a multidisciplinary framework of research quality. The study is aimed at capturing the views of researchers of two groups (SERL-Sweden and ISERN community) with regard to the importance of evaluating research quality.

We conducted a preliminary case study to determine quality alignment among stakeholders within a particular research group. During a series of workshops, we discussed a conceptual model of research quality and employed an electronic form to collect relative values for importance for quality dimensions. We validated the data collection instrument and obtained insights for complementary research.

We further conducted a focus group with experts from the ISERN community. The focus group employed a design alike to our case study research, but focused on the qualitative opinions of participants regarding the nature of research quality. The participants acknowledged the importance of the research quality dimensions. There were, however, concerns regarding (1) the structure of the model, (2) the description of the dimensions, and (3) applicability of the model to research appraisal.

Based on the results from our mixed approach, we provided: 1. description of a method to assess the alignment of stakeholders regarding dimensions of research quality; 2. a revised model of research quality grounded on methodological literature and the suggestions of the experts; and 3. an example for operationalization of the conceptual model to characterize an instrument for research appraisal.

Acknowledgements The work of Jefferson Seide Molléri is supported by the Science Without Borders program, funded by CNPq (National Council for Scientific and Technological Development - Brazil).

Contributors J.M., E.M. and K.P. conceived the idea and planning the overall research. J.M. designed and carried out the pilot case study. J.M. and M.F. designed and conducted the focus group. K.P. and E.M. verified the data collection and synthesis. J.M. analysed the data from both studies. All authors discussed the results and contributed to the final manuscript.

References

1. Roel Wieringa, Nelly Condori-Fernandez, Maya Daneva, Bela Mutschler, and Oscar Pastor. Lessons learned from evaluating a checklist for reporting experimental and observational research. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 157–160. ACM, 2012.
2. Barbara A Kitchenham, O Pearl Brereton, David Budgen, and Zhi Li. An evaluation of quality checklist proposals-a participant-observer case study. In *EASE*, volume 9, page 167, 2009.
3. Martin Host and Per Runeson. Checklists for software engineering case study research. In *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on*, pages 479–481. IEEE, 2007.
4. Pär Mårtensson, Uno Fors, Sven-Bertil Wallin, Udo Zander, and Gunnar H Nilsson. Evaluating research: A multidisciplinary approach to assessing research practice and quality. *Research Policy*, 45(3):593–603, 2016.
5. Sebastian Barney and Claes Wohlin. Software product quality: Ensuring a common goal. In *International Conference on Software Process*, pages 256–267. Springer, 2009.
6. Janet Siegmund, Norbert Siegmund, and Sven Apel. Views on internal and external validity in empirical software engineering. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, volume 1, pages 9–19. IEEE, 2015.
7. Zenda Ofir, Thomas Schwandt, Colleen Duggan, and Robert McLean. Research quality plus (rq+): a holistic approach to evaluating research. 2016.
8. Jerome Kirk, Marc L Miller, and Marc Louis Miller. *Reliability and validity in qualitative research*, volume 1. Sage, 1986.
9. Izak Benbasat and Robert W Zmud. Empirical research in information systems: the practice of relevance. *MIS quarterly*, pages 3–16, 1999.
10. Allen S Lee and Richard L Baskerville. Generalizing generalizability in information systems research. *Information systems research*, 14(3):221–243, 2003.
11. Ray Dawson, Phil Bones, Briony J Oates, Pearl Brereton, Motoei Azuma, and Mary Lou Jackson. Empirical methodologies in software engineering. In *Software Technology and Engineering Practice, 2003. Eleventh Annual International Workshop on*, pages 52–58. IEEE, 2003.

12. Sebastian Barney. *Software Quality Alignment: Evaluation and Understanding*. PhD thesis, Blekinge Institute of Technology, 2011.
13. SL Pfleeger and B Kitchenham. Software quality: the elusive target. *IEEE Software*, pages 12–21, 1996.
14. Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131, 2009.
15. Dimitra Karanatsiou, Yihao Li, Elvira-Maria Arvanitou, Nikolaos Misirlis, and W Eric Wong. A bibliometric assessment of software engineering scholars and institutions (2010–2017). *Journal of Systems and Software*, 147:246–261, 2019.
16. Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.
17. Kai Petersen and Claes Wohlin. Context in industrial software engineering research. In *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 401–404. IEEE Computer Society, 2009.
18. Samireh Jalali and Claes Wohlin. Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 29–38. ACM, 2012.
19. Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *EASE*, volume 8, pages 68–77, 2008.
20. Tony Gorschek, Per Garre, Stig Larsson, and Claes Wohlin. A model for technology transfer in practice. *IEEE software*, 23(6):88–95, 2006.
21. Per Rovegård, Lefteris Angelis, and Claes Wohlin. An empirical study on views of importance of change impact analysis issues. *IEEE Transactions on Software Engineering*, 34(4):516–530, 2008.
22. Ludwik Kuzniarz and Lefteris Angelis. Empirical extension of a classification framework for addressing consistency in model based development. *Information and Software Technology*, 53(3):214–229, 2011.
23. Kai Petersen, Mahvish Khurum, and Lefteris Angelis. Reasons for bottlenecks in very large-scale system of systems development. *Information and Software Technology*, 56(10):1403–1420, 2014.
24. Jyrki Kontio, Laura Lehtola, and Johanna Bragge. Using the focus group method in software engineering: obtaining practitioner and user experiences. In *Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on*, pages 271–280. IEEE, 2004.
25. Workshop on methodological issues in empirical studies with human subjects. `https://isern.iese.de/Portal/mod/page/view.php?id=75`. Accessed: 2018-10-08.
26. Andreas Jedlitschka, Liliana Guzmán, Jessica Jung, Constanza Lampasona, and Silke Steinbach. Empirical practice in software engineering. In *Perspectives on the Future of Software Engineering*, pages 217–233. Springer, 2013.
27. GH Travassos, LFS Silva, RO Spinola, M Kalinowski, and W Chapetta. Tools and facilities for experimentation: Tools and facilities for experimentation: Can we get can we get there. In *Panel presentation on the 11th International Software Engineering Research Network Meeting (ISERN 2003), Italy*, 2003.
28. Victor R Basili and Richard W Selby. Paradigms for experimentation and empirical studies in software engineering. *Reliability Engineering & System Safety*, 32(1-2):171–191, 1991.
29. Martin Ivarsson and Tony Gorschek. A method for evaluating rigor and industrial relevance of technology evaluations. *Empirical Software Engineering*, 16(3):365–395, 2011.
30. Lisa M Given. *The Sage encyclopedia of qualitative research methods*. Sage Publications, 2008.
31. Marten Dorrington Shipman. *The limitations of social research*. Routledge, 2014.
32. Pierre Berthon, Leyland Pitt, Michael Ewing, and Christopher L Carr. Potential research space in mis: A framework for envisioning and evaluating research replication, extension, and generation. *Information Systems Research*, 13(4):416–427, 2002.
33. Anders Mårtensson and Pär Mårtensson. Extending rigor and relevance: Towards credible, contributory and communicable research. In *ECIS*, pages 1325–1333, 2007.
34. Jennifer Mason. *Qualitative researching*. Sage, 2017.
35. Tal Berman. *Conceptual Context*, pages 11–34. Springer International Publishing, Cham, 2017.
36. Roel Wieringa. Towards a unified checklist for empirical research in software engineering: first proposal. 2012.

A Appendices

A.1 Definitions of the quality dimensions according to Mårtensson et al. [4]

#	Criteria	Description
A	Credible	Research that is Coherent, Consistent, Rigorous and Transparent
A1	Rigorous	Research that is Contextual, Internally Valid and Reliable
A1.1	Internally valid	A correct Scientific Method (incl. research design) is used in relation to the Question at Hand and Context, and New Knowledge is Provable.
A1.2	Reliable	The chosen Scientific Method is appropriate for the present Question at Hand and Context, and is documented in a Described Procedure that others could use to reach a similar result in the same Context.
A1.3	Contextual	Existing Knowledge that is relevant for the Context is used, and is presented according to Rules for Description
A2	Consistent	New Knowledge is logically linked to Existing Knowledge and is in accordance with the Scientific Method and Question at Hand
A3	Coherent	Adequate consideration is given to Existing Knowledge in the chosen Context.
A4	Transparent	Relevant New Knowledge in the reporting of research results is included and the process is described in relation to the Question at Hand, Scientific Method and Existing Knowledge
B	Contributory	Research that is Original, Relevant and Generalizable
B1	Original	Research that has an Original Idea, uses an Original Procedure and produces an Original Result
B1.1	Original idea	The Question at Hand has not been asked before in the current Context or is interpreted in a novel way
B1.2	Original procedure	Described Procedure is original in relation to the Question at Hand
B1.3	Original result	New Knowledge is Provable in relation to Existing Knowledge
B2	Relevant	Research that has a Relevant Research Idea, Applicable Result and Current Idea
B2.1	Relevant research idea	Question at Hand is relevant for the current Target Group
B2.2	Applicable result	New knowledge is Beneficial for the current Target Group
B2.3	Current idea	The Question at Hand is in accordance with the current Context
B3	Generalizable	New Knowledge is practically or theoretically useful in Contexts other than the one studied
C	Communicable	The research is consumable, accessible and searchable
C1	Consumable	Research that is Structured, Understandable and Readable
C1.1	Structured	The Research documentation follows the Rules for Description

Continued on next page

Table 4 – *Continued from previous page*

#	Criteria	Description
C1.2	Understandable	The language in the Research documentation is understandable for the Target Group
C1.3	Readable	Correct language in the Research documentation for the Target Group
C2	Accessible	New Knowledge is easily available to the Target Group
C3	Searchable	The documented New Knowledge is structured according to the Rules for Description and easily found by the Target Group
D	Conforming	The research is aligned with regulations, ethical and sustainable.
D1	Aligned with regulations	The Research complies with currently applicable legal aspects of the System of Rules
D2	Ethical	The Research is Morally Justifiable, Open and supports Equal Opportunities
D2.1	Morally justifiable	The Research complies with currently applicable ethical standards as described in the System of Rules
D2.2	Open	The Research demonstrates Transparency with currently applicable ethical standards as described in the System of Rules
D2.3	Equal opportunities	The Research is consistent with equal treatment according to the System of Rules
D3	Sustainable	The Research complies with sustainable development aspects as described in the System of Rules

A.2 Definitions of all concepts in the concept model of research according to Mårtensson et al. [4]

Term	Definition
Actor	A Person initiating and/or performing a Conscious Action
Beneficial	A positive effect of New Knowledge for a Target Group
Conscious action	A process initiated and/or performed by an Actor
Context	An environmental or intellectual setting where the Research takes place and/or is studied, and where Existing Knowledge is valid
Described procedure	A description of how the Research will be performed and documented according to the Rules for Description
Existing knowledge	Knowledge that is built on by the Research, exists in a Context, can be documented in a Source and is expanded with New Knowledge
New knowledge	Knowledge that expands Existing Knowledge, is Provable, and is Beneficial for a Target Group
Person	A human being
Provable	Evidence that the New Knowledge is demonstrable
Question at hand	A research question that is the base for Research
Relationship	A relation between two Conscious Actions showing how those actions interact
Research	A Conscious Action that aims for New Knowledge, emanates from one or several Questions at Hand, studies one or several Contexts, builds upon Existing Knowledge, uses one or several Scientific Methods, is documented in one Described Procedure, requires Transparency and relates to one or several Systems of Rules

Continued on next page

Table 5 – *Continued from previous page*

Term	Definition
Rules for description	Rules describing what a Described Procedure should contain, including its intentions and results. This can differ in regards to Context, Scientific Method, System of Rules, Existing Knowledge and Question at Hand
Scientific method	A described and precise technique used for conducting the Research
Source	Documents, databases or other media that contain Existing Knowledge
System of rules	Legal requirements, regulations, norms and other guidelines that influence how Research should be performed
Target group	Individuals, organizations, enterprises and/or society that benefit from New Knowledge
Transparency	A clear description required by the Research