Information and Software Technology 119 (2020) 106240

Contents lists available at ScienceDirect

Information and Software Technology

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

An empirically evaluated checklist for surveys in software engineering

Jefferson Seide Molléri*, Kai Petersen, Emilia Mendes

BTH - Blekinge Tekniska Högskola, Valhallavägen 1, Karlskrona 37141 Sweden

ARTICLE INFO

Keywords:
Checklist
Assessment
Survey
Methodology

ABSTRACT

Context: Over the past decade Software Engineering research has seen a steady increase in survey-based studies, and there are several guidelines providing support for those willing to carry out surveys. The need for auditing survey research has been raised in the literature. Checklists have been used both to conduct and to assess different types of empirical studies, such as experiments and case studies.

Objective: To operationalize the assessment of survey studies by means of a checklist. To fulfill such goal, we aim to derive a checklist from standards for survey research and further evaluate the appropriateness of the checklist in the context of software engineering research.

Method: We systematically aggregated knowledge from 12 methodological studies supporting survey-based research in software engineering. We identified the key stages of the survey process and its recommended practices through thematic analysis and vote counting. We evaluated the checklist by applying it to existing surveys and analyzed the results. Thereafter, we gathered the feedback of experts (the surveys' authors) on our analysis and used the feedback to improve the survey checklist.

Results: The evaluation provided insights regarding limitations of the checklist in relation to its understanding and objectivity. In particular, 19 of the 38 checklist items were improved according to the feedback received from experts.

Conclusion: The proposed checklist is appropriate for auditing survey reports as well as a support tool to guide ongoing research with regard to the survey design process. A discussion on how to use the checklist and what its implications are for research practice is also provided.

1. Introduction

A survey is a widely deployed research method in the area of Software Engineering (SE) and an increase in its usage has been highlighted by, e.g., Punter et al. [1]. Its purpose is to investigate a population, in order to construct explanatory models [2,3] or to validate knowledge [4,5]. Survey research is often employed when there is a need to study a large set of variables [3] or to perform a retrospective analysis [6]. It may be used to draw conclusions based on both quantitative and qualitative data [7].

Researchers have highlighted various challenges during the survey process. Common challenges are the formulation of questions [8], so to avoid shortcomings (e.g., introducing bias inside questions [3]), and the identification of invalid responses [9]. Other challenges are related to the recruitment of participants, such as how to obtain a sufficient number of responses and how to prevent high drop-out rates [10,11].

There is a value for a systematically constructed checklist for survey research for the following reasons:

– *Need for improvements in survey quality:* The need for improving the standards of conducting and reporting survey-based research, in particular with respect to the definition of the population and sampling strategies is evidenced by Stavru [12]. Furthermore, Stavru pointed out a lack of checklists for auditing surveys in SE which could be of help to both researchers conducting survey research as well as to those evaluating and reviewing the research.
– *Need for checklists:* Checklists are important instruments to support researchers while designing studies and assessing study quality. Software engineering has a set of guidelines for designing surveys (e.g. [13–15]), though no comprehensive checklist has been included in those guidelines.
– *Need for covering multiple survey guidelines in software engineering:* Different guidelines emphasize different aspects of software engineering surveys, e.g. focusing on sampling, or addressing the overall process on a higher level. Hence, by using the survey guidelines for software engineering available at the time of study, a more comprehensive checklist is generated than by looking at individual guidelines.

* Corresponding author.
E-mail address: jefferson.molleri@bth.se (J.S. Molléri).

https://doi.org/10.1016/j.infsof.2019.106240
Received 26 March 2019; Received in revised form 8 November 2019; Accepted 27 November 2019
Available online 3 December 2019
0950-5849/

- *Systematic process of construction:* We should avoid biases when constructing the survey. Biases may be reduced by following a systematic and traceable process for the construction of the checklist.
- *Lack of evaluation:* Only a few guidelines and assessment instruments were evaluated for empirical software engineering (see e.g. [16]). Hence, there is a need to include an evaluation of the checklist.

Motivated by these needs, our contribution is to employ an empirical approach (see Section 3) to **construct and evaluate a checklist for survey-based research in SE**[1].

Looking at other research fields, researchers learned from existing checklists to built new ones [17]. We focused on systematically deriving our checklist from survey research guidelines in software engineering, which were identified in our previous research [18]. In this way, the standards for quality assessment of survey-based studies in our checklist are grounded by the practices recommended by the guidelines to conduct them. As an applied field, it is a goal of Empirical Software Engineering (ESE) to produce high-quality empirical evidence that contributes to the body of knowledge and are relevant to practitioners [19]. It is essential for the maturity of ESE research to operationalize methods and tools for assessing the quality of the produced studies.

The remainder of the paper is structured as follows: Section 2 describes the background and related work. Section 4 details the systematic approach we used to construct the checklist. The evaluation of the checklist in research practice context is presented in Section 5. Section 6 discusses the findings and finally, and Section 7 concludes the paper.

2. Background & related work

We first present existing survey guidelines that are subject-independent or that have been proposed in other fields. After that, we overview survey guidelines and survey assessments in the field of SE. Finally, we look into checklists proposed for assessment of different research methods.

2.1. Survey process and guidelines

Survey as a research method has been established in social research for half a century. It has been employed in several academic fields, such as health care, politics, psychology, and sociology [20]. As a consequence, methodological knowledge on surveys has been published first in these fields.

As the survey research method matured, cross-field guidelines appeared (e.g., [13–15]). These publications aimed to provide methodological support independent from the subject of research. Nevertheless, it is not uncommon for their mentioned practices to focus on the social aspects of research.

Those guidelines [13–15] describe a survey-research process comprising a set of stages, such as question design, sampling, data collection, instrument evaluation, measuring and data analysis [13]. Survey research is acknowledged for being flexible, although the process stages are often conducted sequentially. It is worth mentioning that the survey process is a complete research method (i.e., including planning, execution, analysis, and reporting); the survey data collection instrument is called a questionnaire.

In addition to describing the process, the guidelines also recommend best practices based on desirable attributes for high-quality surveys. Such quality is based on evaluation of the produced evidence (e.g., precision, credibility), ethical issues (e.g., consent, privacy) and mitigating validity threats (e.g., sample error, non-responses) [13,21].

2.2. Survey guidelines in software engineering

The need for specific and tailored guidelines to conduct empirical research in the context of SE has been pointed out, e.g. [19,22,23]. This demand is especially relevant to formal experiments and case studies, due to the popularity of such methods, but also applies to survey-based research. Methodological support for surveys in SE first appeared around the 1990s [16].

Three main guidelines [4,24,25] detail the survey research process in the SE field. They jointly provide a comprehensive structure for the research process, despite differing slightly from each other. Major differences are in relation to the breakdown structure of process stages and the recommended practices provided.

Besides these three main publications, a series of additional studies extend the guidance to particular stages of the survey process. For example, the challenges of identifying the target audience and establishing a sampling frame are discussed in [26–30]. The recommended practices in this set of papers are complementary, although some partially overlap.

Other studies provide lessons learned from carrying out the process in different contexts:

- Punter et al. [1] focus on self-administered online surveys and address issues such as monitoring real-time responses, identifying the reasons for dropouts and encouraging participants to complete a survey instrument;
- Ciolkowski et al. [31] addresses practical issues related to the process itself, such as managing resources and ensuring that deviations do not threaten the completion of the entire process; and
- Cater et al. [32] address replication challenges, such as updating a survey instrument, collecting data and comparing the results.
- Conradi et al. [33] and Ji et al. [34] provide challenges and lessons learned from applying survey research in international contexts, especially regarding invitations and managing responses.
- Torchiano et al. [35] provide lessons learned from experiences in conducting many survey studies, covering topics such as the study design, identifying target audience, sampling, questionnaire design and managing responses.

Additional references for survey-based research are provided in our previous works [16,18].

2.3. Survey assessment

When reviewing existing guidelines (see Section 2.2) we found out that several researchers highlighted the need for an instrument to audit survey research in SE context. This need is further stressed by the lack of reporting of the employed criteria to assess survey research [12].

Stavru's work [12] provides a critical review of surveys in the area of agile software development. In order to carry out the review, Stavru used 21 criteria by which the thoroughness in reporting surveys was assessed. These criteria were extracted from different sources, cf. [4,24,31,36–38]. Note that the method of eliciting the criteria was not detailed.

Stavru also highlighted that the different criteria were not equally important, and rated them on a scale from one to five. The most important criteria that ought to be documented were:

- Sampling frame, method, and size
- Response rate
- Assessment of a survey's trustworthiness
- Survey process
- Conceptual model comprising of the constructs investigated (e.g., variables and their relations)
- Target population
- Questionnaire design

[1] The resulting instrument is further detailed in Appendix A.3. considering the available survey guidelines in software engineering

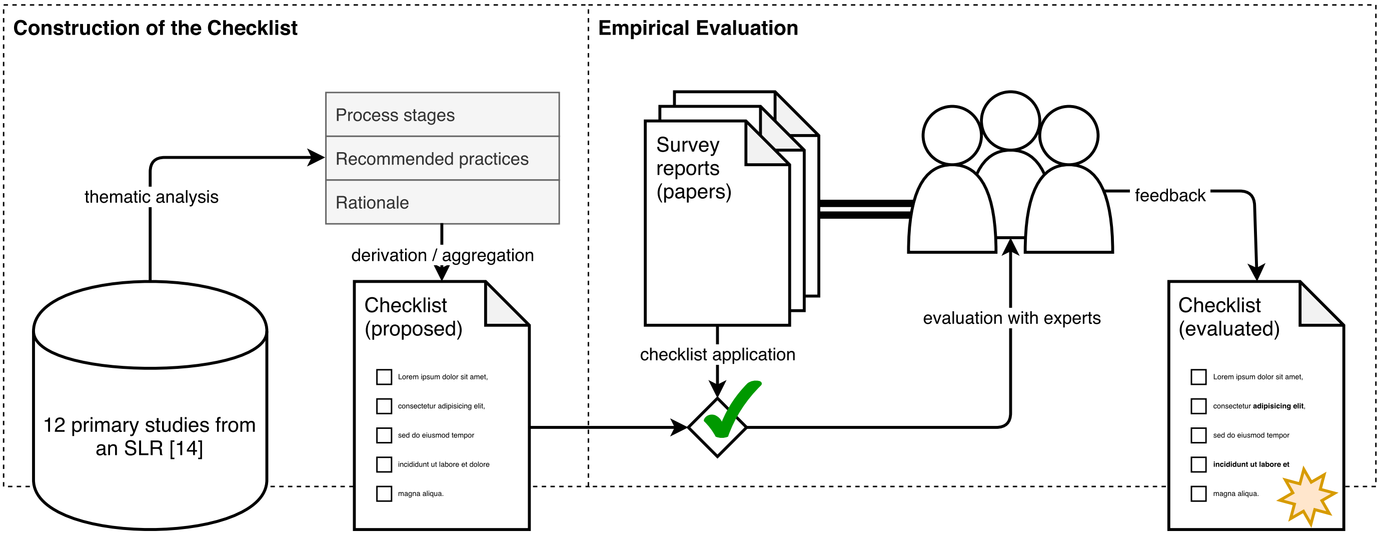


Fig. 1. Overview of our research approach. The first aimed to construct a checklist instrument (see Section 4), and the second step evaluated the checklist (see Section 5).

2.4. Checklists in software engineering

Checklists have been proposed for various research methods with a specific focus on their usage in the SE context. Looking at the ways in which checklists were built, researchers most often based the construction of a new checklist upon existing ones (cf. [39,40]).

As an example, Kitchenham et al. [41] combined two checklists [42,43] to assess experiments and to evaluate whether researchers may use them objectively. Their findings indicate that a larger number of reviewers was needed (eight) to reliably assess studies using their checklist, which could be improved by having researchers conduct reviews in pairs (cf. [41]). Additional checklists proposed for assessing experiments are, e.g., [3,41,44,45].

Höst and Runeson [39] put forward a checklist for case study research, divided according to the research stages, including design, preparation for data & evidence gathering, data analysis, as well as reporting. A validation of the checklist identified that it is appropriated to support researchers conducting case studies, but it is too extensive for reviewing case study reports.

Based on the validation results, Höst and Runeson [39] also created a reduced reviewer's checklist abstracting the original checklist to reduce the number of items to be checked. The 38 items from the researcher's checklist were condensed into 12, most of which synthesizing a list of practices suggested in the original checklist. Additional checklists for reviewing case study research in SE are found in, e.g., [3,46,47].

Wieringa [48] observed that the individual checklists with the same focus differed, which may result in confusions for reviewers. The author highlights the need to find common checklist items across research types as they may share specific aspects. Thus, Wieringa et al. [40] used existing checklists (e.g., [39,44,45,49]) for experiments and case studies as a basis to synthesize an unified checklist. Later, the authors evaluated their checklist by having them used by PhD students and researchers in different research groups, as well as by conference participants.

Stavru's [12] filled a gap in the existing body of knowledge by complementing the set of available checklists with a set of criteria for assessing survey research. No other checklists to assess surveys were identified in our systematic literature search (cf. [18]).

Great emphasis was placed upon (a) basing the checklist on existing literature, and (b) following a systematic approach to eliciting checklist items [12,48]. Thus, our work complements the above-mentioned by deriving and evaluating an assessment checklist grounded in existing guidelines for survey research.

3. Research approach

In order to achieve our main goal to operationalize the assessment of survey studies, we drawn two research questions:

RQ1. What are the standards to be considered for assessing survey-based research in ESE?
RQ2. How appropriate for the research community is a checklist operationalizing such standards?

In the context of our research, standards are concise sets of practices that guide how to conduct and report high-quality surveys. In the context of assessment, standards are mean to be auditable, making it possible for reviewers to verify them. Furthermore, the purpose of adopting a particular standard is herein labeled rationale.

Our empirical research approach comprises two steps, as illustrated in Fig. 1. First, we **constructed a checklist** for operationalize the assessment of survey research in SE. The qualitative method used to derive the checklist was guided by two principles: (a) to identify existing guidance for conducting and reporting survey research; in the context of SE, 12 methodological studies have been considered; and (b) to elicit the recommended practices, the supporting rationale, and related process stages.

The method for systematically deriving the checklist was based on thematic analysis [50]. Vote counting was applied to the themes identified in order to compute the frequency in which they occurred. Further, a co-occurrence was obtained through a relationship matrix relating different categories (e.g., practices versus process stages, and practices versus rationales).

Later, we **evaluated the appropriateness of the checklist** in the context of assessing survey reports. The evaluation process involved two distinct phases: (a) to apply the checklist on a set of published survey reports and register the assessment scores, and (b) to verify the results of this assessment with the survey reports' corresponding authors.

The assessment produced a compliance coefficient for the selected studies in relation to each of the checklist items. We further investigated the authors' feedback in order to understand patterns we identified in the assessment scores. We also collected and addressed suggestions from the experts to improve the checklist instrument.

4. Step 1. construction of the checklist

The first step of our research approach entailed the systematic construction of the checklist. Three sub-contributions are made that ultimately lead to the checklist proposed:

C1. *Consolidation of survey processes and decision points:* We present a consolidated survey process based on existing guidelines. Key decisions points and implications of decision-making are highlighted. For example, a key decision in a survey process is the type of sampling used, which impacts participant recruitment and data analysis. Our checklist has to be adapted depending on the decisions taken.

C2. *Extraction of recommended practices and their mapping to the survey process:* We extracted the recommended practices to be carried out during a survey research process, which were later mapped to the research process identified in C1. Mapping the practices to the main stages aids researchers in the planning of surveys, as it indicates in which process step a practice is executed and where its impact needs to be considered.

C3. *Extraction of rationales for the recommended practices:* The reasons for considering existing survey research practices should be motivated by a rationale, thus making the value of adopting such practices explicit. This is particularly pressing because a survey's cost-effectiveness is an important consideration. Thus, understanding the rationales for the recommended practices supports the cost analysis of a practice and its effectiveness (i.e., the rationale regarding the value a given practice adds to the survey research).

4.1. Method

4.1.1. Research questions

We formulated three research questions corresponding respectively to each of the three contributions stated above, as follows:

RQ1.1. Which stages and key decisions are specified for the survey process (C1)?

RQ1.2. Which practices are suggested and how do they map to the stages of the research process (C2)?

RQ1.3. What is the rationale for conducting the respective recommended practices (C3)?

4.1.2. Study identification and selection

In order to select an appropriate set of primary studies, we used evidence from our previous studies [16,18] identifying methodological papers for survey-based research in SE. A set of three works [4,24,25] provided the main guidelines covering all the stages of a survey research process. We assume that these core papers were likely to hold all the information needed to derive our checklist, i.e., process stages, recommended practices and reasons for their adoption.

Given that only a few papers covering the complete methodology may not cover recommended practices and reasons for their adoption sufficiently, we completed the core set with additional nine additional supporting papers [1,27–29,31,32,51–53] addressing specific stages of survey research such as sampling, instrument design, and validation, recruitment and response management.

4.1.3. Data extraction and analysis

From the set of methodological papers included, we extracted data to address the research questions RQ1.1 to RQ1.3. We employed a thematic analysis process [50] to identify and analyze themes related to three major categories:

- Process stages, e.g., data analysis.
- Recommended practices, e.g., identify reasons for non-responses.
- Rationale attributes, e.g., representativeness.

The thematic analysis process followed the framework proposed by Cruzes and Dyba [50], as follows:

1. **Extract data.** We collected the included studies and aggregated them in a common list using Atlas.ti [54] - a qualitative data analysis software. We also collected bibliographic information for further reference to the included studies.
2. **Code data.** We read all the included studies, identifying segments of the text related to the three main categories and associating them with themes. Besides the main categories, we allowed for the themes to emerge directly from the data using inductive coding. The segments are characterized by a level of granularity of one paragraph, notwithstanding a single paragraph is often associated with more than one code. Paragraphs containing no relevant information were associated with no code.
3. **Translate codes into themes.** The terminology for initial codes is derived from the three guidelines analyzed first (i.e., [4,24,25]). Later, we iteratively improved and updated the initial codes according to the different views presented by the additional sources.
   Successive iterations of the process further refined the theme set, by combining or merging synonyms and aggregating related themes into families (e.g., representativeness is part of the external validity rationale). We also removed duplicates and combined successive occurrences of the same theme in larger segments, thus comprising several paragraphs.
4. **Create a model of higher-order themes.** Later, we identified and analyzed the relationship between themes within the three categories. In particular, we made explicit the recommended practices more frequently related to each process stage and the rationale to adopt them. We computed a co-occurrence coefficient [55] to analyze how frequently two related terms occurred alongside each other. Models describing such relationship are given in Sections 4.3.2 and 4.3.3.
5. **Assess the trustworthiness of the synthesis.** Finally, the identified process stages, recommended practices and rationale are used to answer the research questions. In addition to that, we constructed a checklist instrument grounded on the themes and the relationship among them. We assume that this comprehensive checklist can assess survey research about the relevant practices proposed in the guidelines for SE. Evidence collected from the literature supports this assumption, and we further evaluate the checklist with research practitioners (see Section 5).

4.2. Threats to validity

The empirical construction of the checklist is grounded on a post-positivist research stance, and therefore we discuss four aspects of validity as described by Wohlin et al. [3].

Construct validity. To ensure a similar understanding and to reduce research bias on the thematic analysis, we piloted the coding process between the three authors. The results implied a fair agreement, i.e., on average, 46.5% of the themes were similar, although worded differently. Further, based on the reflections from the pilot study, the first author coded the remaining papers.

The co-occurrence coefficient used in the analysis takes into consideration the position and size of sentences but is prone to non-significant values when comparing themes that differ largely in size [54]. We have partially addressed this potential bias by normalizing the values within the same theme[2]. The normalization mitigated issues of non-signicant values but limited our analysis to comparing themes only within the same category.

Internal validity. Internal validity relates to factors affecting the outcome of the study not accounted for by the researchers. One threat is the bias in interpreting the findings. Hence, at each stage of the research,

[2] Raw co-occurrence coefficients are provided in http://bit.ly/2tRgW2t.

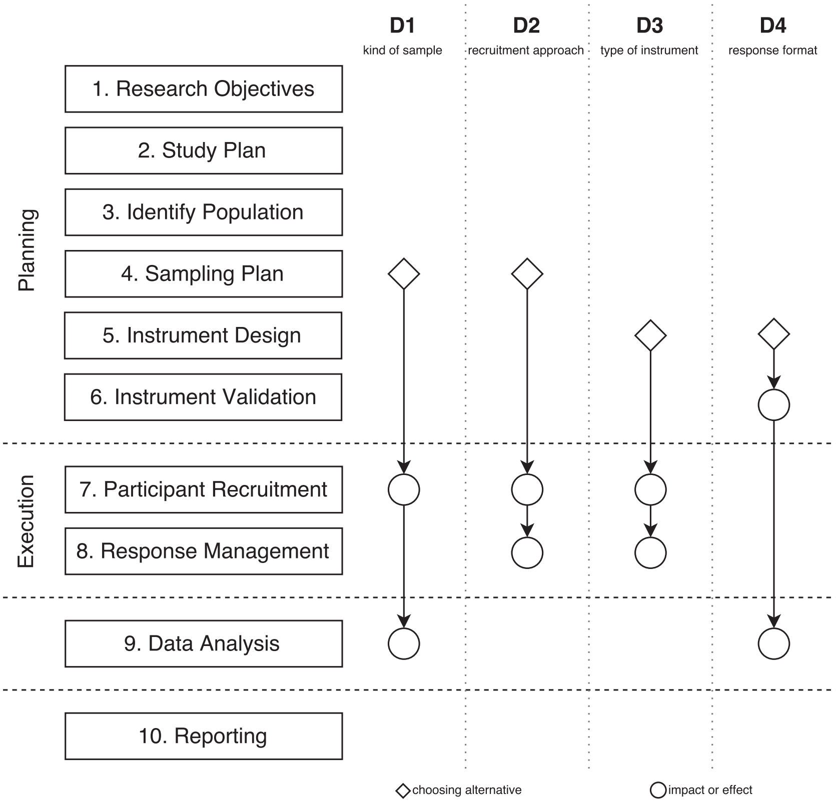


Fig. 2. Model for the survey process, together with the decision points that could impact the research. The diamond symbols (◇,) highlight the stages in which the conditional alternatives should be chosen, and the circles (○) mark the stages affected by those decisions.

the intermediate results were discussed among the researchers (observer triangulation).

For the data extraction, our data set consisted of 12 studies gathered from a previous literature study [18]. We relied mostly on the original study design with regards the search and selection stages, using its reported evidence as primary source for extracting data.

We employed structured reading and coding to analyze the data set, producing themes and higher-level categories. Later, we aggregate the resulting themes and derived the checklist items from them. The first author conducted the data extraction and analysis, further discussing the results with the other co-authors. We trust that this iterative process minimized the judgment bias of a single researcher.

Conclusion validity. One threat to conclusion validity is related to the completeness of the data on which we derived the survey checklist. The additional six papers in our selection (see Section 4.1.2) complemented the study results with ten new practices and one rationale. Those extra themes are related to particular challenges a researcher may face during the process, namely a large sample frame [27,28], managing online surveys [1], and survey replication [32]. We increased the confidence in our results by adding guidelines focused on those particular practices of survey research.

External validity. The resulting themes and frequencies were extracted from relevant methodological guidance for SE. However, we cannot assume that the practices and rationales identified are only important for this field. Moreover, there is the possibility of identifying a valid theme outside of our data set, e.g., a non-selected paper or the practical experiences of a researcher. We, therefore, conducted a evaluation of our proposed checklist with SE researchers, thus investigating its appropriateness and identifying potential improvements (see Section 5).

4.3. Results

The following section is structured according to the three contributions proposed in Section 4.1. Each section is, in turn, broken down into its units of analysis (i.e., decision-making points, process stages, and rationale aggregation).

4.3.1. (C1) Survey research process

Fig. 2 presents an aggregation of the survey research processes described by Kitchenham & Pfleeger [4], Kasunic [24], and Linåker *et al.* [25]. Although the main stages (and terminology) slightly differ among the guidelines, the processes are similar and follow a sequential flow. We adopted the view from [4] to describe the execution phase comprising two stages, one for recruiting participants and one for administering responses.

We also identified key decision-making points during the process, two of which should be addressed during the sampling stage (i.e., D1 and D2) and two others during the instrument design stage (i.e., D3 and D4). Those conditional nodes require researchers to make decisions regarding a survey's research design that can potentially impact the subsequent stages. A decision does not change the sequence of decision points, or exclude any stage that, being part of the path, must also be considered. They can, however, impact how activities in the further stages are performed.

D1. *What kind of sample is selected?* Depending on the strategy for selecting respondents, the researcher can choose a **P) probabilistic** (e.g., random selection) or **NP) non-probabilistic** (e.g., convenience, quota, snowballing) sample. This decision mainly affects the data analysis methods (as probabilistic samples are meant to be generalizable) and recruitment approaches (e.g., random selection) employed.

D2. *How are the participants recruited?* On the one hand, **SS) self-selection** approaches allow for potential respondents to volunteer themselves which may introduce biases in the interpretation of the data.

On the other hand, **PS) personalized selection** (such as invitation letters and more rarely authorization codes) require specific actions for the recruitment and management of the responses.

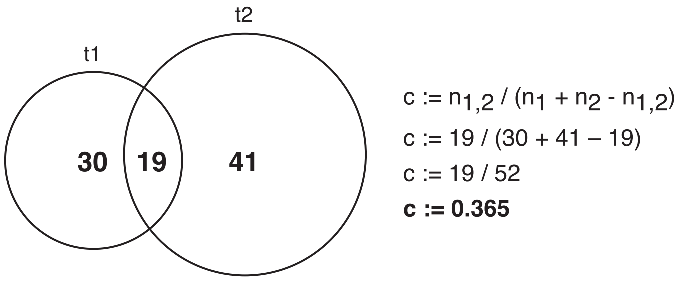


Fig. 3. Example for computation of the co-occurrence coefficient, given that t1 occurs 30 times in the data set, t2 occurs 41 times, and they simultaneously occur 19 times.

This decision-making point is often interdependent of the kind of sample (D1).

D3. *What type of survey instrument is designed?* **SA) Self-administrated** surveys are mainly distributed in the form of Web pages or printed questionnaires, thus the respondents fill out the data themselves.

IA) Interviewer-administrated surveys include face to face or phone interviews where respondents provide the information to a researcher, who records the data. This decision not only drives the instrument design but can also heavily impact the execution stages (i.e., recruitment and response management).

D4. *What response formats are collected?* Question structure types could be **OE) open-ended** and **CE) close-ended**. Open-ended questions are less restrictive allowing for respondents to use their own words, whereas close-ended questions are represented by scales that can be easily quantified. This decision determines the data analysis methods employed, i.e., qualitative or quantitative approaches. Often survey instruments include a mix of both question types, thus requiring both analysis approaches.

4.3.2. (C2) Recommended practices

We identified a list of recommended practices for the survey research process and computed how frequently they occur alongside each other. A co-occurrence coefficient was calculated as follows: $c := n_{1,2}/(n_1 + n_2 - n_{1,2})$, whereas n_1 and n_2 are the vote-counting frequencies of two themes t_1 and t_2 respectively, and $n_{1,2}$ is the joint frequency, i.e., how many times the two themes co-occur. An example of the coefficient computation is given in Fig. 3.

The resulting relationships between recommended practices and process stages are presented in a co-occurrence matrix (see Table 5 in Appendix A.1), where the cells are filled in teal-tones according to the coefficient value (see Section 4.3.2). Darker cells represent a stronger co-occurrence between two themes.

We opted for normalizing the themes in each matrix row since our analysis relates mainly to only comparing themes within the same category. Our normalized coefficient range from 0 to 100, whereas 100 relates to the maximum co-occurrence score in the corresponding group, and 0 corresponds to no co-occurrence. The non-normalized co-occurrence coefficients for each practice in relation to the process stages is available at http://bit.ly/2tRgW2t.

The matrix shows that each stage has a set of practices strongly associated with it, and also which other practices influence that stage. The influence is usually in the form of planning the practice to be carried out, follow-up actions, or consequences of a given decision regarding different practices to adopt (see Section 4.3.1).

A summary of the recommended practices associated to each process stage is given below:

1. *Research objectives:* Survey-based research is motivated by a specific goal. Thus it is important to state the research questions that correspond to such goal. The two main recommendations related to this stage are P1) to limit the scope, as this could impact upon the survey's complexity, and P2) to apply the goal-question-metric (GQM) approach to define its objectives. Moreover, the questionnaire items and collected data should be mapped to the research questions (P44).
2. *Study plan:* The need for designing the survey research is set at the beginning of the process, often along with the research questions [24]. The main suggested practices for this stage are to: P3) investigate related work; P4) define a set of procedures to guide the process; P5) develop a schedule plan for the stakeholders; and P6) start a diary or log book. The study plan should then be iteratively revised during the process, and the updates recorded in the log book (P1). This information is specially required for the reporting stage, at the end of the process.
3. *Identify and characterize the population:* Audience analysis (P7) is often employed to identify and select the characteristics of the population addressed by the research. This task has a strong effect on the sampling stage, in which the sources of sampling (P10) should be defined. Surveys often target potential participants at open databases (P16), but could employ restricted databases (P10) as an alternative or complementary source of sampling. Restricted databases should be investigated prior to the sampling stage.
4. *Sampling plan:* It is often employed in order to sample the population representatively. A sample plan should contain the sources of sampling (P10), units of observation and search unit (P21). The type of sample (P8 and P20) should potentially lead the decision for the data analysis methods employed. Other essential aspects to be considered are the P11) size of the sample and P19) how to manage large samples.

   Additional practices for this stage include to P14) remove the redundant units; P15) apply criteria for selecting the units of observation; P17) plan the retrieval of search units; and P22) partition the population according to the chosen characteristics. Strategies for recruitment (e.g., P8, P18, and P20) are likely to impact the participant selection stage.
5. *Instrument design:* A questionnaire or similar instrument is designed to gather data from the sample representative of the target population. Depending on the choice of distribution, the instruments can be P43) self-administrated, e.g., online forms (P51), or P50) interviewadministrated e.g. interview or phone survey. They can be P29) prototyped, P30) implemented from the sketch, or acquired through P38) commercial tools or P31) reuse.

   Several recommendations to design and present an instrument are provided in the literature, e.g., avoid P33) intrusive and unethical questions, and P57) to lead the respondents; provide P27) a progress indicator, P35) questionnaire navigation, P40) instructions of use, P41) option to resume answering, and mainly P32) ask simple, unambiguous, actual and targeted questions. Responses can assume P46) open-ended or P49) close-ended formats.
6. *Instrument validation:* After design, the ability of the instrument to measure what is intended should be assessed. The most frequently cited approaches for the assessment are P66) piloting, P65) retest, P62) focus groups, and P63) expert or P58) non-expert reviews. Additionally, user-related metrics (e.g., usability, readability, time to respond) can result in improvements to the instrument design (P61). Ancillary documents supporting the recruitment stage should also be reviewed, e.g., cover letter (P60) and thank you letter (P56), likely providing incentives to the respondents (P67).
7. *Participant recruitment:* The strategies to select potential participants are previously defined in the sampling plan stage, such as P24) invitations and authorization codes, P26) self-recruitment, and P13) snowballing. By adopting proper actions and technology support, researchers can even investigate the potential threats to the process related to drop-outs (P68).
8. *Response management:* After distribution of the instrument to the selected participants, it is important to observe the response rate (P64) in order to identify the reasons for non-responses (P25). To ensure

that the expected number of responses is achieved, researchers are likely to send reminders (P70) or to provide rewards for participation (P67).

9. *Data analysis:* Prior to the synthesis, the collected data should be validated (P78) in order to handle incomplete and missing values (P74). Furthermore, qualitative (e.g., P73) or quantitative (e.g., P76) analysis methods can be employed according to the survey's sample and response format. The results should then be P80) presented, P81) interpreted and likely P77) compared to particular subsets of the population. An additional suggestion to ensure their reliability is P79) to have more than one item measuring the same variable.
10. *Reporting:* The main practice related to the reporting phase is to produce an output of the information contained in the process documentation (P6). Ideally, the documentation is to be updated during the survey process, including the data analysis and results' interpretation. Both the related work (P3) and the adopted guidelines (P4) are used as additional information sources for this stage. Finally, it is important to consider the report's intended audience (P7).

The frequency with which the practices occur in the segments of the text is not fully comparable to other practices, i.e., one can not judge a practice as more important if it is mentioned more often. However, the co-occurrence factor can be compared to other instances in the same row to identify in which process stage the practice is more relevant.

Moreover, some recommended practices are mutually exclusive (e.g. P8 probabilistic sample and P20 non-probabilistic sample) requiring the researcher a decision to adopt one of the alternatives. The reasons to adopt a particular practice over another depends on the researchers' conscious decision supported by the guidelines employed.

4.3.3. (C3) Rationales and outcomes

We define a rationale as the motivation to choose a particular practice. They are often described as desirable process attributes (e.g., cost-effectiveness, generalizability) or outcomes of such actions (e.g., minimize or introduce bias). As an example, to achieve generalizability, a researcher should utilize probabilistic sampling (P8) and estimates of the population size (P23).

During the initial step of our coding process, we identified some rationale codes. Later, we combined redundant codes and assembled similar ones into overarching themes. There were similarities between some themes and validity aspects, as described in the Encyclopedia of Survey Research Methods [21]; we labelled them accordingly. Other rationales did not fit this categorization. These were labelled according to targeted aspects, i.e. participant, process, research or result. The process resulted in a set of nine rationale categories, as shown in Fig. 4.

R1. *Conclusion validity:* the actual extent to which conclusions about the investigated relationship are true or correct. Survey-based research employing quantitative analysis methods are prone to significance, effect size, and magnitude factors. Moreover, the reliability or confidence of the results is inherent to the conclusion validity.

R2. *Construct validity:* refers to the interaction between the underlying theory and measurement constructs, i.e., if the variables are actually measuring what they mean to. The main rationale in this category is mensurability, mainly addressed by the instrument validation stage.

R3. *External validity:* the degree to which the results of the survey can be applicable to other scenarios, such as different contexts

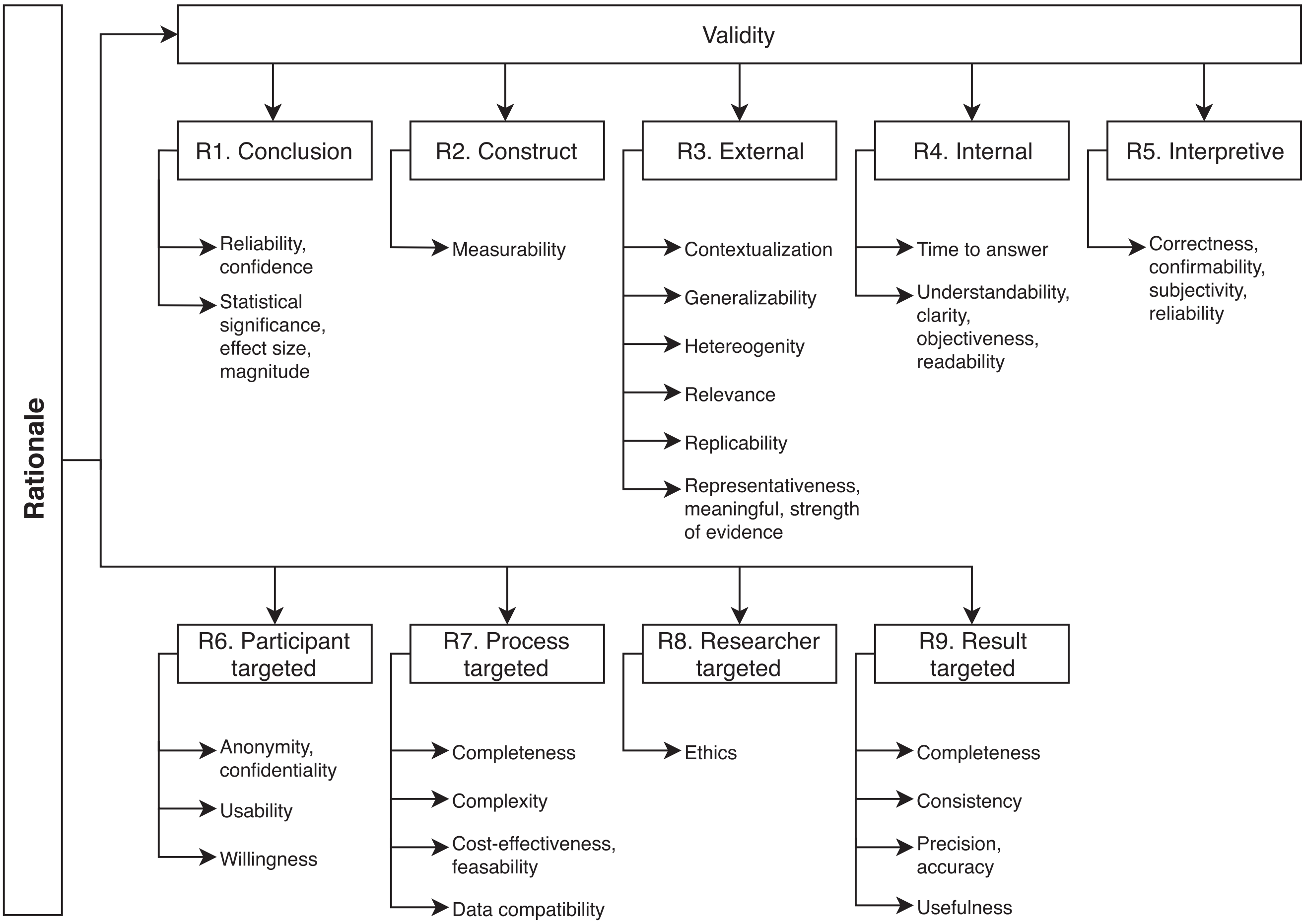


Fig. 4. Rationale Aggregation Tree. Boxes represent the major categories and the related rationales are listed below. The topmost five categories are related to the validity of the research, whereas the bottom four are combined according to targeted aspects, i.e., participants, process, researcher and results.

and strata of the population. Surveys are largely impacted by external validity factors, such as generalizability, replicability, and relevance to practice. Some major factors in this category are whether the sample is representative and heterogeneous to the overall target population.

R4. *Internal validity:* represents an estimate of the degree to which conclusions about the investigated relationships can be drawn based on the measures and the research process. In survey-based research, the rationales in this category are mostly related to the sampling and instrumentation stages, e.g., understandability and time to respond.

R5. *Interpretive validity:* related to the inference of the participants' opinions from the collected responses. Unlike the conclusion validity, the interpretive is more focused on the analysis of the qualitative data. Thus, factors such as correctness, confirmability, and subjectiveness play an important role in interpreting the data.

R6. *Participant-targeted:* additional factors related to the respondents include concerns about anonymity and confidentiality, usability and willingness. Those factors are likely to impact the data quality, as they can positively or negatively influence the participants while answering the survey.

R7. *Process-targeted:* the improvement of the process itself is a target of several rationales. Researchers carrying out surveys should pay special attention to cost-effectiveness, as their decisions are likely to require extra resources. Other practical considerations include the complexity of the instruments and techniques, completeness of the sampling sources, and compatibility of produced data.

R8. *Researcher-targeted:* by providing their opinions, survey respondents trust that the gathered data will be processed responsibly. Thus, it is essential that the researchers are aware of potential ethical issues and their responsibilities regarding the survey process.

R9. *Result-targeted:* one would expect a properly conducted survey process to produce useful results. Data validation tasks are meant to assess data consistency and completeness. Moreover, the precision of the results can be achieved by properly addressing a representative sample of the target population.

Rationales reflect quality attributes of the research process. Moreover, by relating them to the practices can potentially support design decisions. To support such a decision-making process, we also computed a co-occurrence coefficient of the relationships between recommended practices and the rationale (available at https://bit.ly/2UpQqYf). This map makes explicit which practices to adopt to achieve particular standards.

While conducting the research process, it is essential to reflect on the importance of different rationales to be prioritized. As an example, to strengthen the external validity (R3), a researcher should take particular care of strategies for sampling and recruitment of participants, such as using a probabilistic sample (P8), updating the sampling frame based on additional collections from the same search unit (P12), and employing a snowballing technique (P13) to recruit participants.

4.3.4. (C4) Checklist instrument

To derive the checklist, we identified the relevant practices to each process stage according to the co-occurrence index. Further, we aggregated recommended practices addressing a similar topic, e.g. the usability of an online questionnaire is addressed by the set of practices P27, P28, P35, P39, P40, P41, P45, and P51. We then looked up the segments of text coded according to these practices to derive a checklist item. In this way, the formulation of the checklist items closely reflects the practices suggested in the literature.

We also identified the rationale associated with the practices that originated a checklist item (see Section 4.3.3). Our intention by providing the rationale is twofold: (i) to better structure the checklist according to the reasons for adopting the set of practices, and (i) to support the use of the checklist with regard to decisions to be made for employing a set of related practices. Similarly, we identified key decision points that could impact the research (see Section 4.3.1, e.g., D1:P means kind of sample: probabilistic as D1:NP is related to the non-probabilistic alternative).

Later, the first author prepared a draft of the checklist instrument and discussed it together with the other authors. The draft was further refined based on the results from the discussions. Most of the refinements took the form of rephrasing and reordering the checklist items. A few checklist items were integrated to other to improve readability and understandability of the instrument (cf. [39]).

We also identified that most of the checklist items are related to the methodological rigor, and there is a gap about how to report the evidence produced by survey research. Thus, we complemented the reporting section of the checklist with more generic questions selected from the checklist proposed by Dybå and Dingsøyr [56]. Those questions are meant to assess the quality of the evidence produced by empirical studies, regardless of the research method employed.

The checklist resulting from this process had originally 38 items (see pre-evaluation checklist in Appendix A.2). One can notice that several practices are presented within a stage not in accordance to the co-occurrence table (Section 4.3.2). Although some of the practices require early planning, they could eventually be carried out in a later stage. Therefore, the checklist is organized in accordance with the stages in which those actions are more likely to take place.

Several process stages and their recommended practices are subject to decision-making (e.g., the kind of sample, instrument type). The choice should be guided by the motivations (i.e., rationales) and desired outcomes of the process. Due to this decision-making aspect, not all checklist items can be achieved to the same degree at once. We, therefore, rely on the researchers to prioritize the checklist items and hence make trade-offs according to their research goals.

5. Step 2. Evaluation of the checklist

Aligned to the main goal of our study (see Section 3), we summarized the goals, research questions and criteria for our evaluation in Table 1.

The evaluation intended to assess the appropriateness of the checklist and the assessment produced within a research practice context, i.e., with researchers that published survey research papers.

Research question Q1 is focused on the checklist instrument we created in Section 4 and its related completeness and relevance (evaluation criteria EC3 and EC4). Research question Q2 is aimed at the compliance, agreement and fairness (criteria EC1, EC2, EC5) of assessment produced by the checklist. The quantitative criteria EC1 and EC2 were derived from assessing survey articles and comparing the assessment with

Table 1
Evaluation goal, research questions and evaluation criteria.

Goal	Purpose	To evaluate
	Object	a checklist for assessing survey studies in SE
	Issue	with respect to the appropriateness and corresponding evaluation criteria
	Viewpoint	from the researchers' perspective
Research Questions	Q1	Is the checklist operationalizing appropriate standards for assessing survey studies?
	Q2	Is the assessment produced by checklist appropriate from the viewpoint of the researchers?
Evaluation criteria	EC1	**Compliance** of survey reports with the checklist
	EC2	**Agreement** between our assessment and experts' judgement
	EC3	**Completeness** of checklist items
	EC4	**Relevance** for assessing survey studies
	EC5	**Fairness** of the assessment produced

expert judgment, and three qualitative assessment criteria EC3, EC4 and EC5 were obtained only through expert opinion.

5.1. Method

Once the goal has been specified, we selected an evaluation approach for data collection and processing.

The evaluation process employed practices from literature reviews [57] for identifying and selecting candidates, and practices from survey research [24] for recruitment and data collection. Further, we analyzed the data according to both quantitative and qualitative synthesis procedures [50]. The details of the research process are described in the following.

5.1.1. Selection and recruitment

Search strategy. At first, we identified survey-based articles that can be assessed using our checklist. We searched for potential candidates in nine venues (four journals and five conferences) publishing empirical research studies in SE, namely:

TSE	Transactions on Software Engineering
IST	Information and Sofware Technology
EMSE	Empirical Software Engineering Journal
JSERD	Journal of Software Engineering Research and Development
ICSE	International Conference on Software Engineering
SEAA	Euromicro Conference on Software Engineering and Advanced Applications
IWSM-Mensura	International Workshop on Software Measurement
EASE	International Conference on Evaluation and Assessment in Software Engineering
ESEM	International Symposium on Empirical Software Engineering and Measurement

Despite well-established guidelines [4,24], our checklist also incorporate practices mentioned in recent guidelines (e.g. [25,53,53]). Thus, we opted for candidate papers that were published in the 5 most recent years (i.e., from 2012 to 2017), as they are more likely to incorporate such practices. From this database, we identified 3429 potential publications matching these characteristics.

Selection process. We further filtered the papers that mentioned the term “survey” in the title or abstract, thus narrowing the original list down to 177 candidates. We gathered these papers and selected them according to an inclusion criterion: *Does the paper clearly reports survey-based research?* This resulted in 62 included papers.

Recruitment. Later, we invited by e-mail the corresponding authors of the selected papers to participate in our evaluation. Two of the corresponding authors have more than one paper in our candidate list, thus we sent 60 invitations related to 62 resulting papers. The invitation letter presented the goal and the context of the research and also described the assessment procedure (see evaluation procedure, below).

Responses. Three invitation e-mails could not be received with the given e-mail address. Out of the 57 authors who received an e-mail, 22 agreed to participate. One of them consented in assessing two of the papers we asked for and also provided an extra paper which was not part of our dataset. As an incentive to recruiting the participant, we added the additional paper to our list.

5.1.2. Evaluation process

The process to evaluate our checklist consisted of a series of steps. First, we collected 24 referred papers from the corresponding authors that consent to participate in our study, and applied our checklist to assess them. Each of the checklist items was ranked as fully addressed (F), partially addressed (P), not addressed (N), or not applicable (NA). These data relates to metric EC1) compliance.

For each assessed item, we also added notes for possible improvements in the study's documentation, e.g., due to missing information. As an example, in relation to the checklist item 2A, which assess the detailed procedures when designing a survey, we provided the following note to one of the participants: “*The paper cited guidelines to survey research to characterize the sample and recruitment. It is not clear if the method provided in the guidelines are followed thought all the research process.*” In order to preserve the anonymity, we do not report the complete notes here. They were however shared with the corresponding authors.

Later, we provided the corresponding authors with the filled out checklist and our notes so that they could review them. We requested them to provide feedback about our assessment, in order to satisfy EC2) agreement. The respondents pointed out disagreements with our assessment, and refined the assessment scores. We gathered the participants' comments in order to address them individually and identify suggestions for improving our checklist.

Finally, we asked three open-ended questions related to the qualitative criteria, as follows:

(EC3) Do you consider the checklist complete? If not, what should be included?
(EC4) Is there anything you would like to remove, or do you think it is irrelevant?
(EC5) Do you think our assessment by means of this checklist is fair? That is, was our assessment of the paper too rigid or too lenient?

5.1.3. Data analysis

Our analysis considered data collected from three distinct sources:

1. Aggregation of results from our assessment. We aggregated the scores of 24 reported studies from 22 authors that consented having their work assessed by us. We obtained a compliance score (EC1) by the relative percentage of papers that address the checklist items. Both overall scores and the individual scores for each checklist item were analyzed.

2. A review of our assessment. We gathered the participants' scores for each of the checklist items. Based on that, we computed the inter-rater agreement (EC4) between the scores in our assessment and the ones reviewed by the corresponding authors. The inter-rater agreement is expressed in accordance with Cohen's kappa coefficient.

3. Responses to the opinion questions. In order to analyse the review provided by the respondents, we aggregated the participants' answers into a common list[3]. These open-ended answers comprise the respondents own phrasing and reasoning. We read each of the answers and assigned a value in a scale of yes/no/partial. We analysed both information types (i.e., open-ended text and value in the yes/partial/no scale) related to the three qualitative criteria of our evaluation, i.e. *EC3) completeness, EC4) relevance, and EC5) fairness*.

We also addressed the participants' comments and notes from sources 2 and 3, responding to each issue in need of due attention and detailing the actions we took to improve the checklist. In particular, we looked for suggestions for improvement, whether by removing, adding, or rephrasing. The three authors together discussed the comments and contributed to the responses.

5.2. Threats to validity

The evaluation of our checklist followed an interpretivist / constructivist philosophy. We use as basis three aspects of validity for interpretivist research as described by Petersen and Gencel [58].

Credibility. A major threat to validity concerns the ability to assess the constructs with qualitative questions. We relied on the perspective of the participants to judge whether the checklist is complete and relevant to SE research, and whether our assessment using the checklist is fair. In particular, one of the participants questioned whether completeness could be assessed based on opinions.

Another potential threat to validity relates to the interpretation of the findings. In particular, we formulated three open-ended questions

[3] Available at https://goo.gl/XE7wQF.

to collect participants' opinions. Conventionally, open questions are associated to subjective responses, which is likely to constrain the analysis of data. To increase the credibility of our findings, we compared the participants' opinions with the scores resulting from our assessment. These scores are used to identify the practices often not reported or addressed.

The open-ended questions themselves are not bias-free, as they are formulated to extract a positive/negative response. As an example, "Do you consider the checklist complete?" received more positive than negative answers. To decrease this threat, all the three authors interpreted the data resulting from our evaluation study. Further, we discussed any divergences in order to achieve a shared understanding of the qualitative data.

Confirmability. Our great involvement in constructing the checklist is likely to introduce personal biases on our assessment scores. We aimed to mitigate these by building a traceable chain of evidence. First, we assessed the selected papers and recorded notes to support the given scores. We later asked the corresponding authors to review our scores and notes, and to refine any disagreement they identified. We further computed the inter-rater agreement between ours and the participants' scores, resulting in a very strong agreement (k = 0.91, according to weighted Cohen's Kappa [59]).

The inter-rater analysis is commonly employed to assess the agreement between two raters independently. Here, our resulting assessment was shared with the corresponding authors prior to collecting their scores. Specifically, the participants were invited to review the scores they consider unfair and provide us refined ones. We assume that two reviewers using the checklist independently are not likely to achieve such stronger agreement. However, the results from the opinion questions also showed that the corresponding authors judged the assessment as mostly fair (see Section 5.3.5).

Participants provided feedback regarding our proposed checklist and some suggested improvements (see Section 5.3.3). To mitigate bias in understanding participants' feedback, so to act upon each suggestion appropriately, all three authors read and interpreted the comments/suggestions separately. Later a discussion too place to check whether there were divergent views between the three authors. None were found. Given what seemed to be very clear and straightforward comments/suggestions, in addition to similar understanding by all the three authors, we decided not to contact participants again to check whether our interpretation of their feedback was genuine.

However such feedback/suggestions were used as basis for a revised checklist. The first author addressed each change request, and the other two authors corroborated these changes. Furthermore, we aim to further improve the proposed checklist via an additional validation study with the participation of independent reviewers (Section 7).

Transferability. Our selection process aimed to identify a diverse set of survey-based articles, i.e. surveys in different areas and/or surveys of different quality. The sample of papers collected covers a wide range of SE topics, e.g., testing, modeling, and industry practice. From an interpretative perspective, we assume our checklist to be appropriate to assess the variety of topics in SE research and related fields.

The sample of papers we collected were peer-reviewed, so we assume they present a rigorous and sound description from the survey process. This assumption is supported by the results of our assessment, in which the average level of compliance is 65% of the items in our checklist. Thus, our sample is not diverse with regard to the methodological quality of the papers. Besides that, the participation of experienced researchers supports the generalization of our findings by expertise.

5.3. Results

5.3.1. (EC1) Compliance

The resulting scores from **our assessment using the checklist** were aggregated. We further computed compliance coefficient, i.e., the relative amount of papers that addresses the related checklist item, as shown in Table 2. A compliance score of 100% means that all papers were rated "F". The NA ratings are not computed, and each P counts as half of a full score.

The overall compliance with the checklist of selected papers in our sample is 65%. Some of the checklist items presented better compliance, such as the items related to the research objectives (1A to 1C) and three out of four items related to reporting (10B to 10D). One expects any research work, regardless of research method employed, to meet these requirements.

The compliance score could also be interpreted as a record of possible improvements to be taken into consideration for research practice. As an example, the characterization of the target population (checklist item 3A) is seldom reported, thus suggesting a need to foster the adoption of such practice recommended by the existing methodological guidelines.

Three checklist items are fully addressed by all the papers assessed. They cover practices such as incentives to responses (7C), qualitative synthesis (9D) and stratified data analysis (9E). These items are optional, and thus the assessment is rated not applicable (NA) for all the papers that do not employ such strategies. These results imply that researchers applying such strategies are likely to report them explicitly.

Among the checklist groups that are more scarcely addressed are: 2) study plan; 3) identify the population; 6) instrument validation; and 8) response management.

The low compliance scores show that the same kind of information is missing in several assessed studies. This implies that some of the recommended practices proposed by the guidelines are not followed. If we consider that this sample of papers is a good representation of the overall survey-based research in SE, the low-compliance items point out to gaps that should be part of wider discussion so to see if they are relevant in survey research.

5.3.2. (EC2) Agreement

Out of the 22 corresponding authors who agreed to participate, 12 provided us with a feedback pointing out disagreements with our assessment and refining the assessment scores. We compared the respondents' scores to ours via inter-rater agreement. The resulting weighted Cohen's Kappa coefficient k = 0.91 [59], suggesting a very strong level of agreement.

Although we did not employ an independent rating process, the results from the inter-rater agreement are reinforced by the corresponding authors' answers the opinion questions (see Section 5.3.5).

In total, we received 34 comments related to 23 items in our checklist. A subset of our responses are provided in Appendix A.4, and the complete set is available online in https://goo.gl/YDj1XA. The checklist items that required more attention were 5I) additional sources of data collection, and 10A) available materials (questionnaire instrument and ancillary documents). Later, we identified fifteen requests for changes and classified them into clarification, editorial, and structural changes.

- Most of the requests (eleven out of fifteen) were clarification changes intended to improve the understanding and objectiveness of the checklist. In particular, two checklist items (i.e. 4B and 5D) were misinterpreted by the participants, thus requiring some major changes.
- Other three changes were editorial, which include misspelling corrections, updating information or revised wording.
- Finally, only one structural change was needed, splitting one checklist item into two individual ones (now 5D and 5E). It addresses two distinct practices we identified: P32) create simple, unambiguous, actual and targeted questions, and P33) avoid intrusive and unethical questions.

5.3.3. (EC3) Completeness

Most of the respondents (7 out of 12) of our opinion survey agreed that the checklist was complete and included the main aspects of survey-based research. Two participants thought that the checklist was partially complete, and it could be improved by clarifying a few items.

The three remaining participants who did not agree with the checklist completeness, raised issues such as: i) internal and external validity

Table 2
Summary of the combined scores obtained by the papers in our sample. Each row represents a checklist item, and the relative amount of papers (out of 24) ranked as fully addressed (F), partially addressed (P), not addressed (N), or not applicable (NA). The last column computes a compliance score based on how many papers address the related item.

#		N	P	F	NA	Compliance
1. Research Objectives						
1A	Are the research question(s)...	1	0	23	0	95.8%
1B	Is the research context defined?...	1	0	23	0	95.8%
1C	Are the needs for the survey...	1	0	23	0	95.8%
2. Study plan						
2A	Is the survey process supported by guidelines?...	13	2	9	0	41.7%
2B	Is there a reflection on the need to update the research plan?...	19	0	5	0	20.8%
2C	Are the roles and responsibilities...	20	2	2	0	12.5%
3. Identify population						
3A	Is the population characterized...?	13	0	11	0	45.8%
3B	Is the size of the population...	19	1	4	0	18.7%
4. Sampling plan						
4A	Is the kind of sample...defined?	8	5	11	0	56.2%
4B	Is the sample size calculated...	7	1	16	0	68.7%
4C	Are the sources of sampling...	1	2	21	0	91.7%
4D	Are the strategies and criteria to select units...	10	0	14	0	58.3%
5. Instrument design						
5A	Is the type of instrument...defined?	1	1	22	0	93.7%
5B	Is the instrument design process...	5	2	17	0	75%
5C	Are the demographic questions...	2	2	20	0	87.5%
5D	Does particular care is taken to make the questions understandable...?	10	2	12	0	54.2%
5E	Is the number and order of the questions taken in consideration?	16	1	7	0	31.2%
5F	Is there a reflection on the type of responses...for the questions?	4	1	19	0	81.2%
5G	If employing close-ended questions, are the standardized response...	1	3	20	0	89.5%
5H.	Is there a reflection on the adoption of additional sources...	18	2	4	0	20.8%
6. Instrument validation						
6A.	Is the validation process of the survey instrument detailed?...	6	0	18	0	75.0%
6B.	Is the instrument measuring what is intended?...	6	5	13	0	64.6%
6C.	In case of an electronic or online questionnaire, is the usability...	21	2	1	0	8.3%
6D.	Are the results of the instrument validation discussed?...	10	1	12	1	54.3%
7. Participant recruitment						
7A.	Are the strategies to select participants...	0	1	23	0	97.9%
7B.	Are the ancillary documents...	13	4	7	0	37.5%
7C.	If rewards or incentives to respondents are provided...	0	0	2	22	100%
8. Response management						
8A.	Are the responses monitored?...	4	2	18	0	79.2%
8B.	Is there any action to be taken in case of non-responses...?	16	0	5	3	23.8%
9. Data analysis						
9A.	Is the data validated...	16	1	7	0	31.2%
9B.	Is the method for data analysis...	2	2	20	0	87.5%
9C.	If statistical analysis is employed, is the hypothesis testing process...	0	1	15	8	96.9%
9D.	If using qualitative synthesis...	0	0	10	14	100%
9E.	If a stratified sample is defined...	0	0	3	21	100%
10. Reporting						
10A.	Are the instrument and ancillary documents accessible...	5	1	18	0	77%
10B.	Has a discussion of both positive and negative findings...	0	1	23	0	97.9%
10C.	...Are limitations of the study (e.g. threats to validity) discussed?	0	3	21	0	93.7%
10D.	Are the conclusions justified...	0	1	23	0	97.9%
	Mean	11.2	2.04	21.88	2.88	65%

are not completely addressed in relation to the sampling plan and the instrument validation; ii) more details are needed for novice researchers using the checklist; and iii) validating completeness is not possible as an opinion. These comments and change requests were addressed individually in our feedback document (see Appendix A.4).

One participant highlighted their confidence that our method of creating the checklist was grounded in methodological publications, such as [4]. The information regarding the process we used to create the checklist was not provided beforehand, so we assumed that the participant is familiar with such work, thus relating our checklist items to the recommended practice described in Kitchenham's guidelines [4].

5.3.4. (EC4) Relevance

Most of respondents (9 out of 12) considered the checklist items relevant for assessing survey in the context of SE. Three participants mentioned irrelevant checklist items they believed should be removed:

(2C) the checklist item addressing research roles and responsibilities was considered irrelevant for the report, but it could be part of the research plan (2B);

(6A/6C) these two items should be combined, as they both address the instrument validation;

(5H) using additional sources for data collection is optional, therefore if not mentioned in the paper it should be rated NA; and

(7B) to provide ancillary documents (e.g., cover letter, invitation letter) is irrelevant to the research report.

The only issue raised by more than one participant is related to unifying 6A and 6C. The results of our assessment (see Section 5.3.1, below) point out that most of our sample studies are in compliance with 6A (75%), but just a few (8.3%) actually address item 6C. We think that it is important to keep these two aspects separated, thus making explicit the needs for validating the usability of the questionnaire. All the issues abovementioned and the changes requested are discussed in our feedback document (see Appendix A.4).

5.3.5. (EC5) Fairness

Most respondents (9 out of 12) considered our assessment being fair. Two of those also mentioned that despite rigid, the assessment was fair. Another one highlighted the need for instruments that promote rigorous assessment of the research methods. None of the participants described our assessment as completely unfair, although three of them pointed out that items we missed in our assessment (see Section 5.3.3) were limitations to fairness.

Two participants mentioned the lack of information in their reports due to size limitations of the publication. We sympathize with the participants' concern regarding a fair assessment due to the paper's size limitation. However, we stress the importance to provide all the details needed to properly assess the research based on its report. As a recommended practice, researchers are encouraged to make additional information (e.g., research diary, questionnaire instrument, ancillary documents, and other additional material) accessible to the target audience.

6. Discussion

6.1. Checklist usage

In order to assess survey-based research, reviewers can employ the proposed checklist. Prior to assessment, we suggest verifying the availability of research process information (i.e., research report, survey instrument, and ancillary documents). Thereafter, each checklist item should be carefully read and then evaluated with respect to whether the question can be answered and was reflected on in the research report.

Several checklist items comprise two or more nested questions. Those items are intertwined and should not be assessed separately. Moreover, the checklist items can be addressed as partial coverage, due to the higher level of abstraction where answer is likely to be subjective. In such cases, we rely on the reviewers' best judgment regarding the adoption of partial scores (i.e., 0.5).

It is possible to derive a scoring measure based on the checklist marks (e.g., 23 out of 38). However, we do not encourage the simple aggregation of scores in such a way, as it is likely to lead to a loss of assessment information. We suggest reviewers report the reasoning to score each question, thus highlighting the strengths and weaknesses of the assessed survey.

6.2. Implications to research

The objective of our checklist is twofold: first, to audit reported survey-based research; and second, for supporting researchers in making research design decisions and reporting them. Ideally, both the researchers employing the checklist to plan and report their studies and the reviewers assessing the same research should obtain similar scores.

One can derive a tailored checklist instrument focus on the particular needs. As an example, reviewers willing to use the checklist to audit reported survey studies could find our checklist too extensive. Moreover, external reviewers are potentially more interested to assess the evidence provided and quality of report. These aspects are more closely related to the rationale R9) result-targeted and also to the validity aspects of R1) conclusion validity, R3) external validity, and R5) interpretative validity.

By filtering the checklist items according to this set of rationale, we obtain a tailored checklist as illustrated in Table 3. Note that the checklist is more strongly aligned to the process stages S9 Data analysis and S10 Reporting. These stages are more likely to be reported, as well as specific practices such as 1A, 4A, and 4B. We further combined checklist items that relate to the decision-points we identified in Section 4.3.1, as this key practices are likely reported together. The resulting tailored checklist comprised 16 items.

Alternatively, this reflexive checklist can be used to improve the survey process; researchers are encouraged to think and reflect upon the questions they are aiming to use. In particular, trade-offs have to be made. The completeness of the survey as well as the ability to obtain a large and representative sample are desired, but also costly. Thus, as highlighted in the survey guidelines, the research process decisions have to be reflected upon with respect to cost-effectiveness. This is not to say that researchers should aim at minimizing the cost, but rather reflect on what is needed to fulfil the research goals.

Researchers using our checklist are strongly encouraged to report the resulting scores along with their reflections about the checklist itself. We also foster independent evaluations to verify the appropriateness of the checklist to assess survey-based research by the research community.

Finally, the proposed checklist is intended to assess survey-based research in SE, but it has the potential to address different domains' studies (e.g., social sciences). It is important to identify the differences of the survey-based process employed in SE and in other fields, thus evaluating the checklist in a cross-domain study.

6.2.1. Research practice

During the checklist evaluation, we assessed 24 papers reporting survey research. The results of our assessment (Section 5.3.1) point to a list of recommended practices (see Section 4.3.2) that are scarcely addressed. We believe that by communicating these insights to the community, we can encourage researchers to consider the recommended practices in their research. The scarcely addressed practices are:

2. *Study plan:* A research plan or log book (see recommend practice P6) is important to guide the research efforts. This protocol should detail the responsibilities of each stakeholder (P52) and a timetable (P5). The document should be updated as new information becomes known, and ultimately make it accessible by the end of the research.
3. *Identify the population:* Very often the demographics of the participants are described, but scarce information is provided regarding the target population. An audience analysis (P7) is likely to identify and supply these characteristics, and the census of practitioners and institutions could provide estimates of the population size (P23).
5. *Instrument design:* When designing the data collection instrument, one should carefully consider the order (P34) and amount (P36) of questions. Additional sources of data (P69), such as work repositories can provide means to cross-check the results.
6. *Instrument validation:* In general, the reports stated that some kind of validation of the questionnaire (e.g., a pilot) was performed. When employing online surveys, we should ideally check the usability of the questionnaire instrument. A series of practices (see e.g. P27, P28, P35, P39, P40, P41, and P45) can be used as a guideline or checklist to this validation. Furthermore, it is also important to report the improvements made as resulting of the instrument validation (P61).
7. *Participant recruitment:* Besides the need for making the research plan available, it is suggested to provide access to the standardized communication with participants. These documents include the invitation and cover letters (P60) as well as follow-ups and thank you letters (P56).
8. *Response management:* Besides the response rates, it is valuable to report any strategy used to improve responses, such as reminders (P70) or searching for additional databases (P16). These strategies are likely to affect the sample size, thus we should ideally discuss the implications of them for the study validity.

Table 3
An exemplary tailored checklist focused on reviewing aspects of evidence and reporting. These aspects are better represented by the checklist items related to rationale R1, R3, R5, and R9.

Research Objectives	
1A	Are the research objective expressed in measurable terms? E.g. as research questions, or using the goal-question-metric approach.
Identify the population	
3A	Is the population or the survey's target audience characterized (e.g. through audience analysis)?
3B	Is the size of the population stated? If it is not possible to gather this data, are statistic estimates of the population drawn?
Sampling plan and participant recruitment	
4A	Is the kind of sample (i.e. probabilistic, non-probabilistic) defined? Obs. impact for data analysis, its representativeness and/or generalization should be discussed.
4B	Is the sampling process described, and the resulting sample size presented?
4C	Are the sources of sampling (e.g. particular databases or directories, open or restricted) defined? E.g. through a search plan.
4D/7A	Are the strategies to select participants stated and implemented? E.g., through a sampling frame, as well as invitations, authorization codes, self-recruitment, or snowballing.
Response management	
8A	Are the responses monitored? E.g. response rate, non-responsiveness, and drop-out questions. In case of inadequate response rate, the reasons for non-responses and drop-out items were investigated?
Data analysis	
9A	Is the data validated prior to analysis? E.g. through checking inconsistent, incomplete and missing values.
9B	Is the method for data analysis specified? Are the steps of the analysis process described? Are they suitable for the response formats collected?
5G/9C	If statistical analysis is employed, is the hypothesis testing process documented and the standardized responses (i.e. nominal, ordinal, interval or ratio) stated? Appropriate scales should be assigned according to the mapped variables.
5C/9E	Are the demographic questions formulated according to the audience? If a stratified sample is defined, are the data analysed according to demographics? Are there meaningful comparisons drawn from them?
Reporting	
10A	Are the instrument and ancillary documents accessible (e.g. URL link, external reference, appendix) to readers? If not, are the reasons for that discussed and convincing? If data resulting from the survey were disclosure, were anonymity and confidentiality of data discussed?
10B	Has a discussion of both positive and negative findings been demonstrated? Are the discussion addressing the research question(s) or hypothesis? Does the discussion take into consideration the generalization of the findings?
10C	Are the results of the assessment checklist reported? Are limitations of the study (e.g. threats to validity) discussed?
10D	Are the conclusions justified by the results? Furthermore, are the implications and potential use of the results discussed?

9. *Data analysis:* Before the analysis, researchers are encouraged to validate the data (P78) and check for inconsistent, incomplete or missing information (P74). There are several strategies to deal with missing data, from discarding to the imputation based on statistical models. Ideally, the implications due to employing such approaches should be described.

6.3. Comparison with related work

We previously identified related studies providing checklists for assessing empirical research in SE (see Section 2.4). The existing checklists in SE target mostly experiments [41,44,45] and case study research [40]. Our proposed checklist is intended to address the issues of survey-based research and its specific stages, e.g., sampling, instrument design, and recruitment.

Our resulting checklist can be comparable to Stavru's set of criteria to assess the thoroughness of surveys. Similar to their work, we also systematically derived our instrument from the literature. Stavru's criteria were extracted from a set of six methodological papers, three of them from other fields (i.e. information and communication, and management). The remaining three methodological papers are also part of our data set. Due to this different primary sources, their criteria differs from ours in the following aspects:

Our checklist covers a set of practices grounded on a more comprehensive review of the SE literature. We identified a set of 81 recommended practices that resulted in a checklist containing 39 items in comparison to Stavru's 20 criteria. Most of their criteria have a corresponding item in our checklist (as shown in Table 4), although they are often phrased differently. Four of Stavru's criteria differ conceptually from our checklist:

– **Conceptual model:** our checklist assesses the objects that are investigated and their relation with the questionnaire items (P44), although a model of the variables and their relationships are not explicitly covered;
– **Provisions for securing trustworthiness:** our checklist assesses the adoption of additional sources for data validation, one of the options being supporting literature;
– **Response burden:** the time to answer the questions is a motivation (i.e. related to rationale R4) to be considered when designing the questionnaire instrument; and
– **Assessment of trustworthiness:** our checklist does not explicitly cover statistical approaches for error analysis; rather, a more general question addresses a well-documented and suitable data analysis for the responses collected.

In addition to that, three criteria from Stavru's work do not have a correspondent item in our checklist:

– **Sponsorship:** two of the methodological papers [4,24] mention sponsors with regards to the cover letter (P60), and this information should be provided when appropriate;
– **Execution time:** the period of administering a questionnaire could be informed in the invitation (P24) [4]. It is rather essential to monitor the response rates (P71) and take actions (P70) to remind participants; and
– **Responses:** reporting the number of respondents seems intuitive, however we did not identify any explicit recommendation for doing that. Rather, a discussion on the findings motivated by the interpretation of data (P81) is recommended.

Different from Stavru's, our checklist items are not weighted, as the relative importance of each practice depends on the survey process,

Table 4
Comparison between Stavru's work [12] and our constructed checklist. The first column lists the criteria for thoroughness, whereas the second column lists the corresponding items in our checklist. Items marked with an asterisk are conceptually different in our checklist. The rightmost column points out the related recommended practices we identify by our systematic approach.

Stavru's criteria for thoroughness	Our evaluated checklist	
	Checklist questions	Related practices
Survey definition		
Objectives	1A	P2
Sponsorship	–	–
Survey method	2A	P4
Survey design		
Conceptual model	1B, 6B (*)	P44
Target population	3A	P7, P15, P21
Sampling frame	4C, 4D	P10
Sampling method	4A	P8, P20
Sample size	4B	P11
Questionnaire design	5F, 5G	P34, P36, P46, P49
Provisions for securing trustworthiness	5I (*)	P4, P69
Survey implementation		
Questionnaire evaluation	6A	P54, P62, P65, P66
Questionnaire	10A	–
Survey execution		
Media	7A	P24, P26
Execution time	–	–
Response burden	5F (*)	P36, P60
Follow-up procedures	8B	P70
Responses	–	–
Response rate	8A	P71
Survey analysis and packaging		
Assessment of trustworthiness	9B, 9C (*)	P72, P76
Limitations	10C	–

and researchers are encouraged to reflect on the key decision points. Finally, we evaluated our checklist with experienced researchers. The evaluation identifies room for improvement, and we provide an updated post-evaluation version in Appendix A.3. More generic checklists [48,56] share some similarities with our proposed checklist. Similar questions are mostly related to the research objectives and reporting stages. This is expected since some actions (e.g., formulate research questions, define the scope, discuss the results and limitations of research) are inherent to all empirical research.

We considered the question formulation of the generic checklists when phrasing our proposed instrument. Researchers who used those similar checklists are likely to recognize the overlaps. This could be beneficial, as they can employ identical reasoning when assessing the common items. Similar questions provide the opportunity to compare the checklists or the studies assessed through them.

7. Conclusions

In this paper, we described a two-step empirical approach to operationalize the assessment of survey-based research in software engineering. The motivation for such work is grounded in the increased usage and the need to assess the quality of survey-based research in SE. In a previous study, we identified several guidelines for conducting and reporting survey studies, but without any method or tool to assess their quality.

In our first step, we used a thematic analysis approach to elicit standards for survey-based research. We identified process stages, recommended practices and reasons for adopting them. A set of 12 methodological studies provided qualitative data that resulted in a set of themes and co-occurrence frequencies. Based on this, we built the proposed checklist comprising 38 questions organized by the survey process stages.

Our systematic approach shows that it is possible to derive an assessment instrument from well-established research methodologies. As the checklist is based on recommended practices, a research can also employ it to guide through the process, supporting key decision-making steps. Moreover, the rationale provides a way to filter a set of standards aligned to a desired outcome or purpose.

In the second step, we empirically evaluated the assessment produced by the checklist with experts. We provided the experts with an assessment produced by applying the checklist to their own reported surveys. The sample of analyzed surveys reports comply with 65% of the standards in our checklist. This result is strongly agreed by the corresponding authors. Overall, our the assessment was considered rigorous and fair and the instrument was evaluated as complete.

Experts' feedback also provided us with suggestions to improve the understandability and the objectivity of the checklist. We reflected upon the suggestions and conducted changes in our proposed checklist (the post-evaluation checklist is available in Appendix A.3). We acknowledge this evaluation as a preliminary step for further application and evolution of the instrument.

Our two-steps approach demonstrated how to operationalize the assessment of survey studies by means of constructing and evaluating a checklist. We believe that the empirical software engineering community can benefit from our checklist. It can be a valuable asset for both researchers conducting and reporting survey-based studies, and for reviewers auditing survey reports. Furthermore, an approach similar to ours can be used to operationalize the assessment of different research methods, such as action research, interviews and observational studies.

As future work, we have identified two potential developments from our work:

- With our evaluation, we illustrated how the proposed checklist can be used to assess the quality of reported survey studies. However, the evaluation is based on our own assessment of the studies and the subjective opinions of the participants. It is important to investigate the potential benefits of using the checklist in a more realistic context. To achieve this objective, we plan to conduct a validation study with independent reviewers.
- Similar checklists exist for assessing the quality of different research methods (e.g., case study, experiment) in the context of SE. It would be beneficial to the SE research community to understand the core similarities among these instruments. The similarities presumably represent quality standards for ESE research regardless of the methodology adopted. We intend to compare our proposed checklist with other assessment tools (e.g. [40,41,44,45]) and to provide a shared instrument aligned to views of research quality.

Author Contribution

All authors contributed to conceiving the idea and planning the research. J.M carried out the construction and evaluation of the checklist. K.P. and E.M. verified the data extraction and synthesis of guidelines. All authors discussed the results and contributed to the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

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).

Appendix A

A1. Co-occurrence tables

Table 5
Co-occurrence matrix of recommended practices (rows) according to the process stages (columns). Each cell in the table has a normalized coefficient, i.e., the highest co-occurrence value in each row is assigned a value 100 whereas the lowest value is 0. Cell shading illustrates the strength of the normalized co-occurrence coefficient, ranging from white (0) to teal (100).

		Process stages									
	Recommended practices	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
P1	Avoid too many objectives	100									
P2	Goal-question-metric (GQM)	100		20		20					
P3	Identify related studies	93	57	36	100	64	21	29	21	71	71
P4	Systematically follow procedures / guidelines	80	20	20	100	67	33	20	33	60	80
P5	Produce a schedule or timetable		100			54			15		
P6	Keep a diary or log book	38	23	8	31	15	50		23	35	100
P7	Audience analysis	11	11	100	44	22		11	6	33	72
P8	Kind of sample: probabilistic		7	10	100			3			
P9	Restricted databases, directories, groups, and subjects			63	100						
P10	Source of sampling (SoS)			58	100			25			
P11	Calculate sample size		8	10	100		2	11	5	3	
P12	Update sampling frame based on additional collection of the same unit		38		100			38			
P13	Snowballing selection			25	100			38			
P14	Remove duplicates / redundant units		15	10	100			40			
P15	Criteria for selecting the units of observation	10	50	10	100			40			
P16	Open databases, social networks, digital libraries		16	38	100			44			
P17	Search plan		63	13	100			50			
P18	Randomize sample		9	30	100			27		6	
P19	Increase sample size (or large-samples)			12	100	24		29	12		
P20	Kind of sample: non-probabilistic		4	13	100	4		15		15	
P21	Define units of observation/analysis and search unit	16	16	23	100			26	10	13	
P22	Create subgroups of the population (strata)		4	11	100	11		25	2	16	5
P23	Estimate the population size (through statistics)			18	100					55	
P24	Recruitment approach: personalized selection		14		100	43	21	86	46	7	
P25	Identify reasons for non-responses		63		100	50	44	63	88	13	
P26	Recruitment approach: self-selection		20		100	33		53	20		53
P27	Completion measure or progress indicator					100					
P28	Non-mandatory answers					100					

(continued on next page)

Table 5 (*continued*)

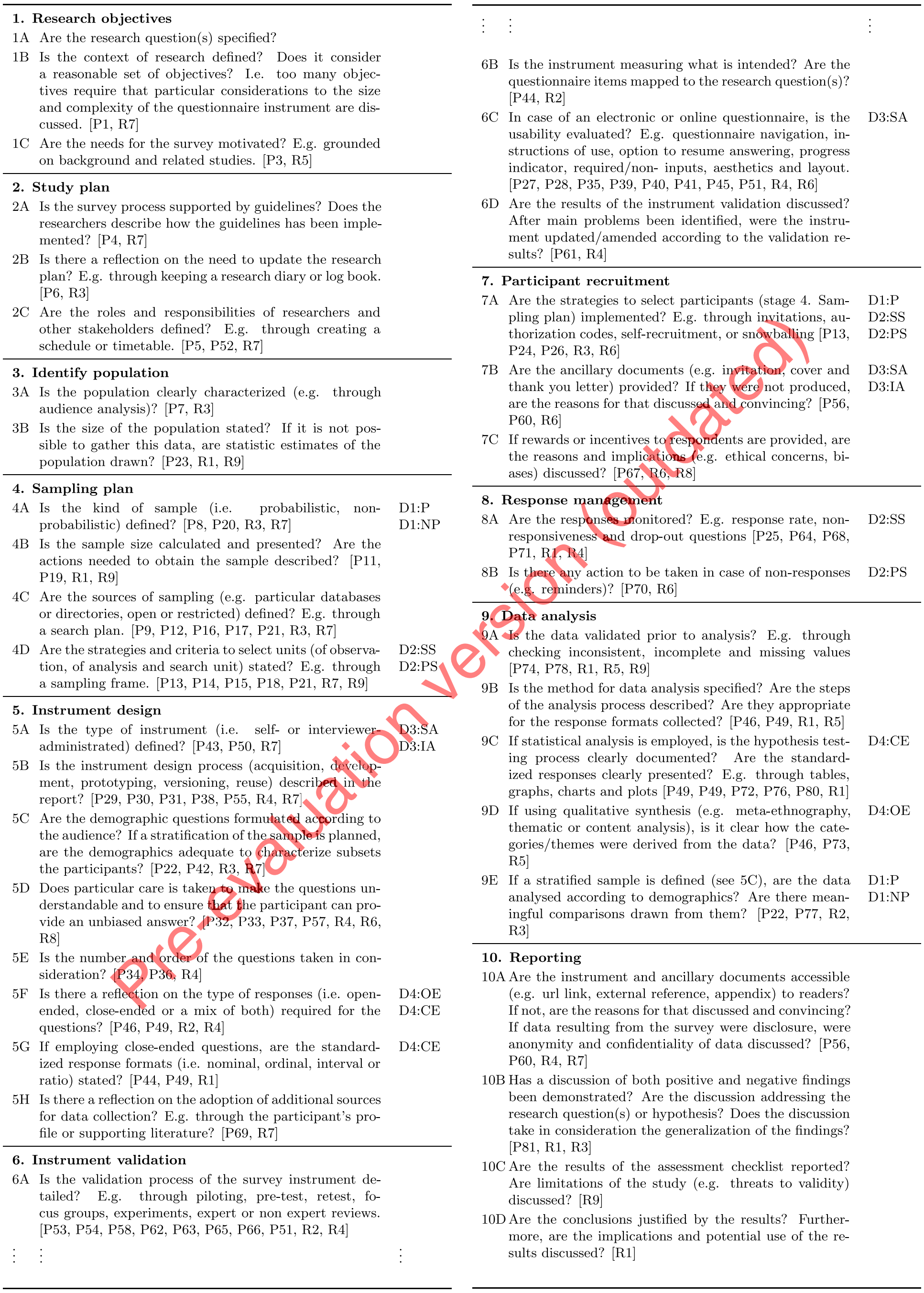
P29	Prototype survey instrument					100					
P30	Self-develop survey instrument					100					
P31	Reuse survey instrument		40			100					
P32	Create simple, unambiguous, actual and targeted questions	1	3	1		100	19		6	1	
P33	Avoid intrusive and unethical questions		20			100			13		
P34	Prioritize questions					100			15		
P35	Questionnaire navigation					100			15		
P36	Prune questions		39			100	33		11		
P37	Define terminology	30	7	7		100	33		13		
P38	Commercial survey instrument		25			100	50		17		
P39	Questionnaire structure and format				6	100	37		9	17	
P40	Survey instructions		9			100	63	13	19	6	
P41	Allow resume answering					100	100		50		
P42	Investigate demographic information			29	82	100		35	6	24	
P43	Type of instrument: self-administrated/printed questionnaires					100		17	42	17	
P44	Map items to research objectives	97	18	24	11	100			3	29	8
P45	Additional technological support		9		61	100	26	22	35	22	
P46	Response format: open questions (unrestrictive)	7				100			4	48	
P47	Error checking					100	67		22	56	
P48	Improve response rates	28			28	100	22	33	33		28
P49	Response format: close-ended questions (standarize responses)	3			20	100	27		20	83	
P50	Type of instrument: interviewer-administrated					100		40	60	60	
P51	Online questionnaires	7	27	27	40	100	67	30	70	60	33
P52	Define roles and responsibilities	85	54	15		100	23		8	62	92
P53	Validation approach: compare instrument to a “gold standard”						100				
P54	Validation approach: inter-rater agreement measure						100				
P55	Alternative versions of the instrument						100	11			
P56	Ancillary document: thank you or follow-up letter					63	100				
P57	Avoid leading participants					67	100				
P58	Validation approach: non-expert reviews					8	100			12	
P59	Assessment through empirical research				10		100		10	10	
P60	Ancillary document: cover letter		21			63	100	37	11		

(continued on next page)

Table 5 (*continued*)

P61	Amend or update the instrument	14	18	9		64	100	18	14	9	
P62	Validation approach: focus group, discussion, reasoning	50		25	19	56	100			31	
P63	Validation approach: expert review	62		15	38	77	100	15		54	
P64	Monitor response rate		57		79	29	100	14	93	14	
P65	Validation approach: test-retest				50		100			75	
P66	Validation approach: pilot or pre-test	9	9	4	48	61	100		13	78	
P67	Provide rewards or incentives to respondents	31	23			77	100	38	54		54
P68	Identify reasons for drop-outs							100	80		
P69	Additional sources of data collection								100		
P70	Send reminders		27			82		45	100		
P71	Monitor respondents in real-time					50		75	100	75	
P72	Discriminant analysis									100	
P73	Categorize qualitative data (i.e. coding)	29	8	21	8	46		8	13	100	
P74	Handling inconsistent, incomplete or missing data		15		15	10		10	25	100	
P75	Data triangulation	83							33	100	
P76	Statistical analysis	7	7	10	62	7			3	100	14
P77	Data comparison		29		29	24			6	100	24
P78	Data validation								11	100	39
P79	More than one question measuring the same aspect				23	69	77		15	100	38
P80	Data representation: tables, graphs, charts and plots								7	100	48
P81	Interpret the resulting data	13	20			60			20	100	73

A2. Survey assessment checklist (pre-evaluation)



1. Research objectives

1A Are the research question(s) specified?

1B Is the context of research defined? Does it consider a reasonable set of objectives? I.e. too many objectives require that particular considerations to the size and complexity of the questionnaire instrument are discussed. [P1, R7]

1C Are the needs for the survey motivated? E.g. grounded on background and related studies. [P3, R5]

2. Study plan

2A Is the survey process supported by guidelines? Does the researchers describe how the guidelines has been implemented? [P4, R7]

2B Is there a reflection on the need to update the research plan? E.g. through keeping a research diary or log book. [P6, R3]

2C Are the roles and responsibilities of researchers and other stakeholders defined? E.g. through creating a schedule or timetable. [P5, P52, R7]

3. Identify population

3A Is the population clearly characterized (e.g. through audience analysis)? [P7, R3]

3B Is the size of the population stated? If it is not possible to gather this data, are statistic estimates of the population drawn? [P23, R1, R9]

4. Sampling plan

4A Is the kind of sample (i.e. probabilistic, non-probabilistic) defined? [P8, P20, R3, R7] — D1:P
D1:NP

4B Is the sample size calculated and presented? Are the actions needed to obtain the sample described? [P11, P19, R1, R9]

4C Are the sources of sampling (e.g. particular databases or directories, open or restricted) defined? E.g. through a search plan. [P9, P12, P16, P17, P21, R3, R7]

4D Are the strategies and criteria to select units (of observation, of analysis and search unit) stated? E.g. through a sampling frame. [P13, P14, P15, P18, P21, R7, R9] — D2:SS
D2:PS

5. Instrument design

5A Is the type of instrument (i.e. self- or interviewer-administrated) defined? [P43, P50, R7] — D3:SA
D3:IA

5B Is the instrument design process (acquisition, development, prototyping, versioning, reuse) described in the report? [P29, P30, P31, P38, P55, R4, R7]

5C Are the demographic questions formulated according to the audience? If a stratification of the sample is planned, are the demographics adequate to characterize subsets the participants? [P22, P42, R3, R7]

5D Does particular care is taken to make the questions understandable and to ensure that the participant can provide an unbiased answer? [P32, P33, P37, P57, R4, R6, R8]

5E Is the number and order of the questions taken in consideration? [P34, P36, R4]

5F Is there a reflection on the type of responses (i.e. open-ended, close-ended or a mix of both) required for the questions? [P46, P49, R2, R4] — D4:OE
D4:CE

5G If employing close-ended questions, are the standardized response formats (i.e. nominal, ordinal, interval or ratio) stated? [P44, P49, R1] — D4:CE

5H Is there a reflection on the adoption of additional sources for data collection? E.g. through the participant's profile or supporting literature? [P69, R7]

6. Instrument validation

6A Is the validation process of the survey instrument detailed? E.g. through piloting, pre-test, retest, focus groups, experiments, expert or non expert reviews. [P53, P54, P58, P62, P63, P65, P66, P51, R2, R4]

⋮ ⋮ ⋮

⋮ ⋮ ⋮

6B Is the instrument measuring what is intended? Are the questionnaire items mapped to the research question(s)? [P44, R2]

6C In case of an electronic or online questionnaire, is the usability evaluated? E.g. questionnaire navigation, instructions of use, option to resume answering, progress indicator, required/non- inputs, aesthetics and layout. [P27, P28, P35, P39, P40, P41, P45, P51, R4, R6] — D3:SA

6D Are the results of the instrument validation discussed? After main problems been identified, were the instrument updated/amended according to the validation results? [P61, R4]

7. Participant recruitment

7A Are the strategies to select participants (stage 4. Sampling plan) implemented? E.g. through invitations, authorization codes, self-recruitment, or snowballing [P13, P24, P26, R3, R6] — D1:P
D2:SS
D2:PS

7B Are the ancillary documents (e.g. invitation, cover and thank you letter) provided? If they were not produced, are the reasons for that discussed and convincing? [P56, P60, R6] — D3:SA
D3:IA

7C If rewards or incentives to respondents are provided, are the reasons and implications (e.g. ethical concerns, biases) discussed? [P67, R6, R8]

8. Response management

8A Are the responses monitored? E.g. response rate, non-responsiveness and drop-out questions [P25, P64, P68, P71, R1, R4] — D2:SS

8B Is there any action to be taken in case of non-responses (e.g. reminders)? [P70, R6] — D2:PS

9. Data analysis

9A Is the data validated prior to analysis? E.g. through checking inconsistent, incomplete and missing values [P74, P78, R1, R5, R9]

9B Is the method for data analysis specified? Are the steps of the analysis process described? Are they appropriate for the response formats collected? [P46, P49, R1, R5]

9C If statistical analysis is employed, is the hypothesis testing process clearly documented? Are the standardized responses clearly presented? E.g. through tables, graphs, charts and plots [P49, P49, P72, P76, P80, R1] — D4:CE

9D If using qualitative synthesis (e.g. meta-ethnography, thematic or content analysis), is it clear how the categories/themes were derived from the data? [P46, P73, R5] — D4:OE

9E If a stratified sample is defined (see 5C), are the data analysed according to demographics? Are there meaningful comparisons drawn from them? [P22, P77, R2, R3] — D1:P
D1:NP

10. Reporting

10A Are the instrument and ancillary documents accessible (e.g. url link, external reference, appendix) to readers? If not, are the reasons for that discussed and convincing? If data resulting from the survey were disclosure, were anonymity and confidentiality of data discussed? [P56, P60, R4, R7]

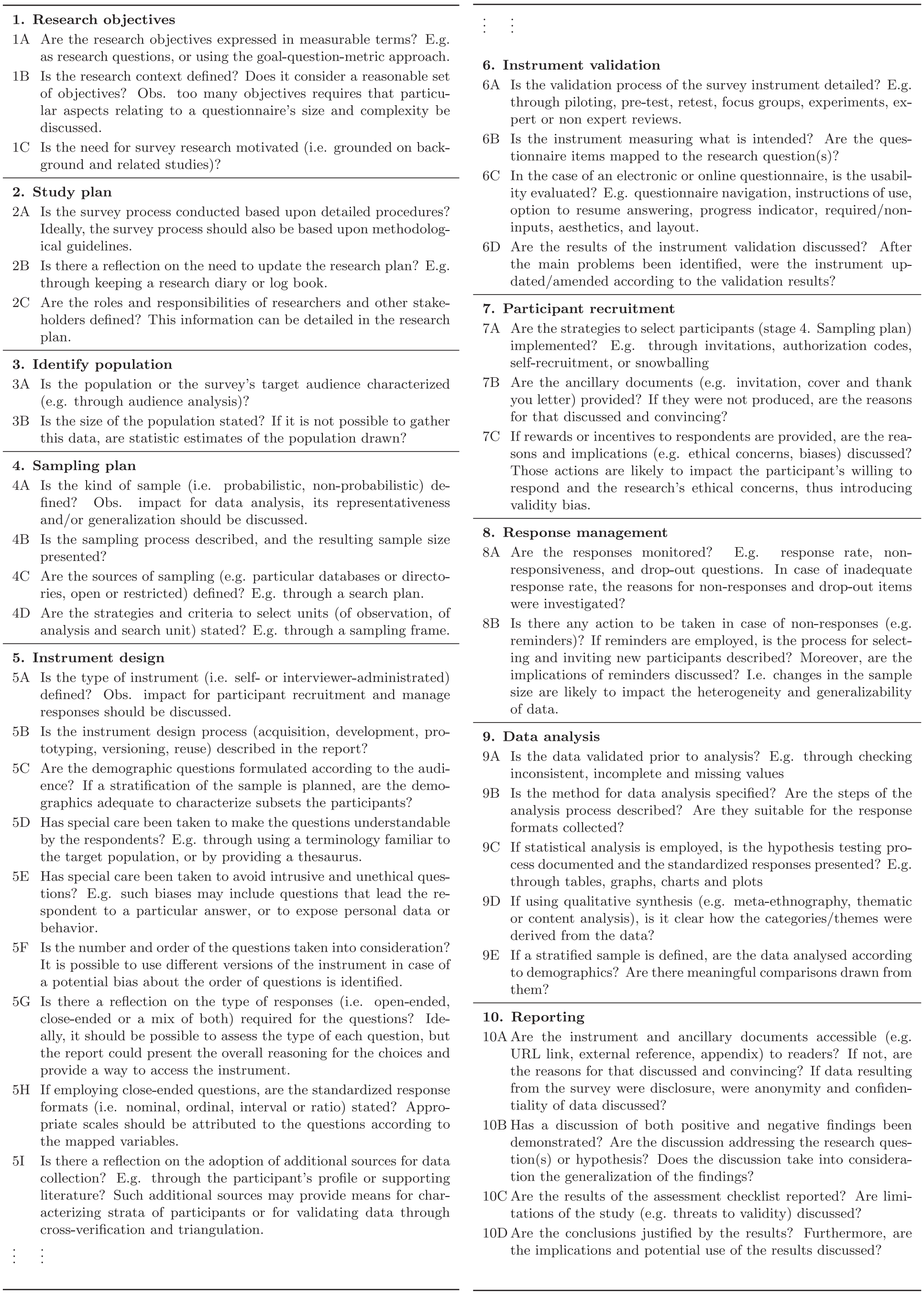
10B Has a discussion of both positive and negative findings been demonstrated? Are the discussion addressing the research question(s) or hypothesis? Does the discussion take in consideration the generalization of the findings? [P81, R1, R3]

10C Are the results of the assessment checklist reported? Are limitations of the study (e.g. threats to validity) discussed? [R9]

10D Are the conclusions justified by the results? Furthermore, are the implications and potential use of the results discussed? [R1]

Fig. 5. Survey assessment checklist proposed. This pre-evaluated version is later improved and updated (see Appendix A.3).

A3. Survey assessment checklist (post-evaluation)



1. Research objectives

1A Are the research objectives expressed in measurable terms? E.g. as research questions, or using the goal-question-metric approach.

1B Is the research context defined? Does it consider a reasonable set of objectives? Obs. too many objectives requires that particular aspects relating to a questionnaire's size and complexity be discussed.

1C Is the need for survey research motivated (i.e. grounded on background and related studies)?

2. Study plan

2A Is the survey process conducted based upon detailed procedures? Ideally, the survey process should also be based upon methodological guidelines.

2B Is there a reflection on the need to update the research plan? E.g. through keeping a research diary or log book.

2C Are the roles and responsibilities of researchers and other stakeholders defined? This information can be detailed in the research plan.

3. Identify population

3A Is the population or the survey's target audience characterized (e.g. through audience analysis)?

3B Is the size of the population stated? If it is not possible to gather this data, are statistic estimates of the population drawn?

4. Sampling plan

4A Is the kind of sample (i.e. probabilistic, non-probabilistic) defined? Obs. impact for data analysis, its representativeness and/or generalization should be discussed.

4B Is the sampling process described, and the resulting sample size presented?

4C Are the sources of sampling (e.g. particular databases or directories, open or restricted) defined? E.g. through a search plan.

4D Are the strategies and criteria to select units (of observation, of analysis and search unit) stated? E.g. through a sampling frame.

5. Instrument design

5A Is the type of instrument (i.e. self- or interviewer-administrated) defined? Obs. impact for participant recruitment and manage responses should be discussed.

5B Is the instrument design process (acquisition, development, prototyping, versioning, reuse) described in the report?

5C Are the demographic questions formulated according to the audience? If a stratification of the sample is planned, are the demographics adequate to characterize subsets the participants?

5D Has special care been taken to make the questions understandable by the respondents? E.g. through using a terminology familiar to the target population, or by providing a thesaurus.

5E Has special care been taken to avoid intrusive and unethical questions? E.g. such biases may include questions that lead the respondent to a particular answer, or to expose personal data or behavior.

5F Is the number and order of the questions taken into consideration? It is possible to use different versions of the instrument in case of a potential bias about the order of questions is identified.

5G Is there a reflection on the type of responses (i.e. open-ended, close-ended or a mix of both) required for the questions? Ideally, it should be possible to assess the type of each question, but the report could present the overall reasoning for the choices and provide a way to access the instrument.

5H If employing close-ended questions, are the standardized response formats (i.e. nominal, ordinal, interval or ratio) stated? Appropriate scales should be attributed to the questions according to the mapped variables.

5I Is there a reflection on the adoption of additional sources for data collection? E.g. through the participant's profile or supporting literature? Such additional sources may provide means for characterizing strata of participants or for validating data through cross-verification and triangulation.

⋮ ⋮

⋮ ⋮

6. Instrument validation

6A Is the validation process of the survey instrument detailed? E.g. through piloting, pre-test, retest, focus groups, experiments, expert or non expert reviews.

6B Is the instrument measuring what is intended? Are the questionnaire items mapped to the research question(s)?

6C In the case of an electronic or online questionnaire, is the usability evaluated? E.g. questionnaire navigation, instructions of use, option to resume answering, progress indicator, required/non-inputs, aesthetics, and layout.

6D Are the results of the instrument validation discussed? After the main problems been identified, were the instrument updated/amended according to the validation results?

7. Participant recruitment

7A Are the strategies to select participants (stage 4. Sampling plan) implemented? E.g. through invitations, authorization codes, self-recruitment, or snowballing

7B Are the ancillary documents (e.g. invitation, cover and thank you letter) provided? If they were not produced, are the reasons for that discussed and convincing?

7C If rewards or incentives to respondents are provided, are the reasons and implications (e.g. ethical concerns, biases) discussed? Those actions are likely to impact the participant's willing to respond and the research's ethical concerns, thus introducing validity bias.

8. Response management

8A Are the responses monitored? E.g. response rate, non-responsiveness, and drop-out questions. In case of inadequate response rate, the reasons for non-responses and drop-out items were investigated?

8B Is there any action to be taken in case of non-responses (e.g. reminders)? If reminders are employed, is the process for selecting and inviting new participants described? Moreover, are the implications of reminders discussed? I.e. changes in the sample size are likely to impact the heterogeneity and generalizability of data.

9. Data analysis

9A Is the data validated prior to analysis? E.g. through checking inconsistent, incomplete and missing values

9B Is the method for data analysis specified? Are the steps of the analysis process described? Are they suitable for the response formats collected?

9C If statistical analysis is employed, is the hypothesis testing process documented and the standardized responses presented? E.g. through tables, graphs, charts and plots

9D If using qualitative synthesis (e.g. meta-ethnography, thematic or content analysis), is it clear how the categories/themes were derived from the data?

9E If a stratified sample is defined, are the data analysed according to demographics? Are there meaningful comparisons drawn from them?

10. Reporting

10A Are the instrument and ancillary documents accessible (e.g. URL link, external reference, appendix) to readers? If not, are the reasons for that discussed and convincing? If data resulting from the survey were disclosure, were anonymity and confidentiality of data discussed?

10B Has a discussion of both positive and negative findings been demonstrated? Are the discussion addressing the research question(s) or hypothesis? Does the discussion take into consideration the generalization of the findings?

10C Are the results of the assessment checklist reported? Are limitations of the study (e.g. threats to validity) discussed?

10D Are the conclusions justified by the results? Furthermore, are the implications and potential use of the results discussed?

Fig. 6. Survey assessment checklist after evaluation (see Section 5). A digital version of the checklist is available at https://tinyurl.com/se-survey-checklist.

A4. Suggestions to improve the checklist (excerpt)

Here we present a sample of the feedback provided by the participants (i.e., corresponding authors) of our evaluation in the professional context (Section 5). The comments are listed according to the checklist item they are related to. For each comment, we present our responses and actions we took to address the mentioned topics. A complete list detailing all the comments is provided at https://goo.gl/jNXx7U.

(1B) Two comments regarding the understandability of this checklist item:

(i) *What type of context and limitations should be described? Many questions would benefit from having a more detailed guide along with the checklist.*; and

(ii) *The term ”limitations of scope” may be misleading. Reading quickly I first thought that you referred to whether the study scope has some limitations to be able to answer RQs (related to study validity).*

Response: We agree that term ”limitations of scope” can leave room for interpretations. It could also be misleading, as limitations are often described as threats to the validity of a study.

Action: To improve the checklist understanding, we rephrased item 1B, as follows: “Is the research context defined? Does it consider a reasonable set of research objectives? Obs. too many objectives requires that particular aspects relating to an instrument’s questionnaire’s size and complexity be discussed.”.

(2) Two comments regarding the description of this checklist item:

(i) *The sub-questions for me do not address the main question of whether a survey is appropriate. 2A-C are more about what is reported, rather than whether survey is the right method. That for me is more about whether other approaches were considered etc. Roles and responsibilities I would generally not note in a paper.*; and

(ii) *(...) I am not convinced that the three questions that are included in this category would be enough to assess if a survey study research design is appropriate to address its research aims (as it is stated in the question). The fact that guidelines are followed, a research diary is kept and responsibilities are defined does not guarantee that the research design is appropriate to answer the research questions. (...)*

Response: We agree with the authors that the description of checklist item 2 does not match what is assessed in sub-items 2A-2C. These sub-items assess whether the study plan is accessible and complete, instead of “appropriate”.

In order to assess whether the survey design is appropriate to address its research objectives, we designed a specific question (see 6B), that assess if the questionnaire items are related to the research questions described in the study plan.

The recommended practices for the checklist item 2 are: (i) to provide a survey plan document; (ii) to report the guidelines used; and (iii) to detail the responsibilities of each researcher. These aspects are important to allow for the study to be reviewed and replicated.

Action: We updated phase2′s description to match what is assessed by items 2A-2C: “Study plan: Is the survey design accessible and complete?”.

(2C) Two comments regarding the relevance of this checklist item:

(i) *Is this information really relevant for the report? I think this information would fits better into a protocol or research plan than in the report itself.*; and

(ii) *It sounds irrelevant (e.g., how relevant it is for assessing the survey itself to know timetables)?*

Response: We agree with the authors that a schedule or timetable is not relevant for assessing the quality of the survey. However, these artifacts can provide means to assess the roles and responsibilities of researchers in the survey process [24]. As suggested by one of the participants, the information regarding the roles and responsibilities of each researcher could be provided in a survey plan document. We highlight here the need to make this document accessible to reviewers [32].

Action: We removed the references to schedule and timetables in the item 2C, instead stating of that “This information can be detailed in the research plan (see item 2B)”.

(5H) Three comments regarding the understandability of this checklist item:

(i) *I didn’t understand it.*;

(ii) *It makes no sense for me. If man don’t mention another source of information it means there is not. Why do you assume that there is another unmentioned source?*; and

(iii) *I don’t fully follow this question. Do you mean data triangulation so findings from the survey are triangulated with other data?*

Response: The authors pointed out a very important understanding issue. Our proposed checklist accounts for the need to discuss if additional sources are required, e.g., to characterize stratas of the participants, or to cross validate the data related to the investigated phenomenon from multiple sources [24,25]. As an example, after the survey, the findings can be compared to other sources, such as personal profile information or related work.

Action: We added a note on 5H making explicit the reasons to adopt additional sources of data collection, as follows: “Such additional sources may provide means for characterizing strata of participants or for validating data through cross-verification and triangulation”.

(10A) *Is data disclosure / open data also a criterion? I think it should be as people should be pushed in the general direction of open science to foster reproducibility.*

Response: The author suggests that, besides the ancillary documents we already mentioned in the checklist, the resulting data from the survey is also make available. We see the value on that, but acknowledge potential implications due to anonymity and confidentiality that should be taken into consideration. Therefore, we rely on the judgment of researchers conducting the survey to provide a discussion on such aspect.

Action: We added a note on 10A making explicit that the data disclosure can be provided and thus should be discussed. The new item is “If data resulting from the survey are disclosure, does a consideration about the anonymity and confidentiality of data is discussed?”.

(10A) Two additional comments regarding applicability of this checklist item:

(i) *I can’t imagine including things like invitations, thank you notes, etc. My ethics approval process requires so much documentation, I would need a separate 10 pages just for all of the information I provide to participants.* ; and

(ii) *not all materials can be added to an appendix especially with paper length limitation.*

Response: We agree with the author that limitations due the publication size are likely to constrain the amount of information provided in a paper. As an alternative, the additional documents mentioned in 10A could compose a ”survey research package”, available as a web reference provided in the paper [32].

References

[1] T. Punter, M. Ciolkowski, B. Freimut, I. John, Conducting on-line surveys in software engineering, in: Empirical Software Engineering, 2003. ISESE 2003. Proceedings. 2003 International Symposium on, IEEE, 2003, pp. 80–88.

[2] E.R. Babbie, Survey research methods., Wadsworth, 1973.

[3] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, Experimentation in Software Engineering, Springer, 2012.

[4] B.A. Kitchenham, S.L. Pfleeger, Principles of survey research: parts 1 – 6, ACM SIGSOFT Softw. Eng. Notes 26–28(2001–2003).

[5] B. Kitchenham, S. Linkman, D. Law, Desmet: a methodology for evaluating software engineering methods and tools, Comput. Control Eng. J. 8 (3) (1997) 120–126.

[6] S.L. Pfleeger, Experimental design and analysis in software engineering, Annal. Softw. Eng. 1 (1) (1995) 219–253.

[7] C.W. Dawson, Projects in Computing and Information Systems: a Student's Guide, Pearson Education, 2005.

[8] M. Torchiano, F. Ricca, Six reasons for rejecting an industrial survey paper, in: Conducting Empirical Studies in Industry (CESI), 2013 1st International Workshop on, IEEE, 2013, pp. 21–26.

[9] C. Yang, P. Liang, P. Avgeriou, A survey on software architectural assumptions, J. Syst. Softw. 113 (2016) 362–380.

[10] R. Akbar, M.F. Hassan, A. Abdullah, A framework of software process tailoring for small and medium size IT companies, in: Computer &amp; Information Science (ICCIS), 2012&nbsp;International Conference on, IEEE, 2012, pp. 914–918.

[11] M. Galster, D. Tofan, Exploring web advertising to attract industry professionals for software engineering surveys, in: Proceedings of the 2nd International Workshop on Conducting Empirical Studies in Industry, ACM, 2014, pp. 5–8.

[12] S. Stavru, A critical examination of recent industrial surveys on agile method usage, J. Syst. Softw. 94 (2014) 87–97.

[13] F.J. Fowler Jr, Survey Research Methods, Sage publications, 2013.

[14] P.L. Alreck, R.B. Settle, The Survey Research Handbook, McGraw-Hill, 1994.

[15] A. Fink, The Survey Handbook, 1, Sage, 2003.

[16] J.S. Molléri, K. Petersen, E. Mendes, Cerse-catalog for empirical research in software engineering: a systematic mapping study, Inf. Softw. Technol. 105 (2019) 117–149.

[17] A. Tong, P. Sainsbury, J. Craig, Consolidated criteria for reporting qualitative research (coreq): a 32-item checklist for interviews and focus groups, Int. J. Qual. Health Care 19 (6) (2007) 349–357.

[18] J.S. Molléri, K. Petersen, E. Mendes, Survey guidelines in software engineering: an annotated review, in: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2016, Ciudad Real, Spain, September 8–9, 2016, 2016, pp. 58:1–58:6.

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

[20] J.M. Converse, Survey Research in the United States: Roots and Emergence 1890–1960, Routledge, 2017.

[21] P.J. Lavrakas, Encyclopedia of Survey Research Methods, Sage Publications, 2008.

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

[23] C. Wohlin, Is there a future for empirical software engineering? in: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2016, p. 1.

[24] M. Kasunic, Designing an effective survey, Technical Report, CMU/SEI-2005-HB-004. Software Engineering Institute, Carnegie Mellon University., 2005.

[25] J. Linåker, S.M. Sulaman, R. Maiani de Mello, M. Höst, Runeson, P. (2015). Guidelines for conducting surveys in software engineering. Technical report #5366801, Sweden: Lund University, https://lup.lub.lu.se/search/publication/5366801.

[26] R.M. de Mello, G.H. Travassos, Would sociable software engineers observe better? in: Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on, IEEE, 2013, pp. 279–282.

[27] R.M. de Mello, P.C. da Silva, G.H. Travassos, Sampling improvement in software engineering surveys, in: Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2014, p. 13.

[28] R.M. de Mello, P.C. da Silva, P. Runeson, G.H. Travassos, Towards a framework to support large scale sampling in software engineering surveys, in: Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2014, p. 48.

[29] R.M. de Mello, P.C. Da Silva, G.H. Travassos, Investigating probabilistic sampling approaches for large-scale surveys in software engineering, J. Softw. Eng. Res. Devel. 3 (1) (2015) 1–26.

[30] R.M. de Mello, G.H. Travassos, Surveys in software engineering: identifying representative samples, in: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2016, p. 55.

[31] M. Ciolkowski, O. Laitenberger, S. Vegas, S. Biffl, Practical Experiences in the Design and Conduct of Surveys in Empirical Software Engineering, Springer, 2003.

[32] A. Cater-Steel, M. Toleman, T. Rout, Addressing the challenges of replications of surveys in software engineering research, in: Empirical Software Engineering, 2005. 2005 International Symposium on, IEEE, 2005, pp. 10–pp.

[33] R. Conradi, J. Li, O.P.N. Slyngstad, V.B. Kampenes, C. Bunse, M. Morisio, M. Torchiano, Reflections on conducting an international survey of software engineering, in: Empirical Software Engineering, 2005. 2005 International Symposium on, IEEE, 2005, pp. 10–pp.

[34] J. Ji, J. Li, R. Conradi, C. Liu, J. Ma, W. Chen, Some lessons learned in conducting software engineering surveys in China, in: Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2008, pp. 168–177.

[35] M. Torchiano, D.M. Fernández, G.H. Travassos, R.M. de Mello, Lessons learnt in conducting survey research, in: Proceedings of the 5th International Workshop on Conducting Empirical Studies in Industry, IEEE Press, 2017, pp. 33–39.

[36] A. Pinsonneault, K. Kraemer, Survey research methodology in management information systems: an assessment, J. Manag. Inf. Syst. 10 (2) (1993) 75–105.

[37] A.K. Shenton, Strategies for ensuring trustworthiness in qualitative research projects, Educat. Inf. 22 (2) (2004) 63–75.

[38] M.K. Malhotra, V. Grover, An assessment of survey research in pom: from constructs to theory, J. Oper. Manag. 16 (4) (1998) 407–425.

[39] M. Höst, P. Runeson, Checklists for software engineering case study research, in: Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, ESEM 2007, September 20–21, 2007, Madrid, Spain, 2007, pp. 479–481.

[40] R. Wieringa, N. Condori-Fernández, M. Daneva, B. Mutschler, O. Pastor, Lessons learned from evaluating a checklist for reporting experimental and observational research, in: 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '12, Lund, Sweden - September 19, - 20, 2012, 2012, pp. 157–160.

[41] B. Kitchenham, D.I.K. Sjøberg, P. Brereton, D. Budgen, T. Dybå, M. Höst, D. Pfahl, P. Runeson, Can we evaluate the quality of software engineering experiments? in: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM 2010, 16–17 September 2010, Bolzano/Bozen, Italy, 2010.

[42] T. Dybå, T. Dingsøyr, Strength of evidence in systematic reviews in software engineering, in: Proceedings of the Second International Symposium on Empirical Software Engineering and Measurement, ESEM 2008, October 9–10, 2008, Kaiserslautern, Germany, 2008, pp. 178–187.

[43] B.A. Kitchenham, O.P. Brereton, D. Budgen, Z. Li, An evaluation of quality checklist proposals-a participant-observer case study., in: EASE, 9, 2009, p. 167.

[44] A. Jedlitschka, D. Pfahl, Reporting guidelines for controlled experiments in software engineering, in: 2005 International Symposium on Empirical Software Engineering (ISESE 2005), 17–18 November 2005, Noosa Heads, Australia, 2005, pp. 95–104.

[45] A. Jedlitschka, M. Ciolkowski, D. Pfahl, Reporting experiments in software engineering, in: Guide to advanced empirical software engineering, Springer, 2008, pp. 201–228.

[46] B. Kitchenham, L. Pickard, S.L. Pfleeger, Case studies for method and tool evaluation, IEEE Softw. 12 (4) (1995) 52–62.

[47] D.E. Perry, S.E. Sim, S.M. Easterbrook, Case studies for software engineers, in: Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on, IEEE, 2004, pp. 736–738.

[48] R. Wieringa, Towards a unified checklist for empirical research in software engineering: first proposal, in: 16th International Conference on Evaluation &amp; Assessment in Software Engineering, EASE 2012, Ciudad Real, Spain, May 14–15, 2012. Proceedings, 2012, pp. 161–165.

[49] K.F. Schulz, D.G. Altman, D. Moher, Consort 2010 statement: updated guidelines for reporting parallel group randomised trials, BMC Med. 8 (1) (2010) 18.

[50] D.S. Cruzes, T. Dyba, Recommended steps for thematic synthesis in software engineering, in: Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on, IEEE, 2011, pp. 275–284.

[51] B.A. Kitchenham, S.L. Pfleeger, Personal opinion surveys, in: Guide to Advanced Empirical Software Engineering, Springer, 2008, pp. 63–92.

[52] B.A. Kitchenham, D. Budgen, P. Brereton, Surveys, in: Evidence-Based Software Engineering and Systematic Reviews, CRC Press, 2015, pp. 234–242.

[53] R.M. de Mello, G.H. Travassos, Characterizing sampling frames in software engineering surveys, in: Proc. 12th Workshop on Experimental Software Engineering (ESELAW), 2015.

[54] S. Friese, Atlas. ti 7 user manual, Berlin: ATLAS. ti Scientific Software Development GmbH (2012).

[55] R.B. Contreras, Examining the context in qualitative analysis: the role of the co-occurrence tool in atlas. ti, Newsletter 2011 (2011) 2.

[56] T. Dybå, T. Dingsøyr, Empirical studies of agile software development: a systematic review, Inf. Softw. Technol. 50 (9) (2008) 833–859.

[57] B.A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, 4, CRC Press, 2015.

[58] K. Petersen, C. Gencel, Worldviews, research methods, and their relationship to validity in empirical software engineering research, in: Proceedings of the 2013 Joint Conference of the 23Nd International Workshop on Software Measurement (IWSM) and the 8th International Conference on Software Process and Product Measurement, in: IWSM-MENSURA '13, IEEE Computer Society, Washington, DC, USA, 2013, pp. 81–89, doi:10.1109/IWSM-Mensura.2013.22.

[59] J. Cohen, Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit., Psychol. Bull. 70 (4) (1968) 213.