Guidelines for Conducting Surveys in Software Engineering

Linåker, Johan; Sulaman, Sardar Muhammad; Maiani de Mello, Rafael; Höst, Martin

2015

Guidelines for Conducting Surveys in Software Engineering
v. 1.1

Johan Linåker, Sardar Muhammad Sulaman, Martin Höst,
Software Engineering Research Group,
Department of Computer Science,
Lund University, Sweden

Rafael Maiani de Mello
Experimental Software Engineering Group
Systems Engineering and Computer Science Department
COPPE/UFRJ [1]

June 2, 2015

[1]Supported by CAPES Scholarship PDSE-99999.014261/2013-08 (Brazil)

Contents

Chapter 1

Introduction

This report is the result of a Ph.D. course on survey methodology in software engineering research at the Department of Computer Science, Lund University, Sweden. All chapters, except for this introduction, are written by Sardar Muhammad Sulaman, Johan Linåker, Rafael Maiani de Mello.

The basic idea of survey methodology is to collect information from a group of people by sampling individuals from a large population. Examples of surveys are found in daily life in several situations, such as election polls, markets surveys, etc. and there is a large amount of literature on the general methodology. The intention of this report is to present guidelines that are adapted to software engineering research. Software engineering research is multi-disciplinary and a survey in this area, of course, resembles surveys in other areas when it comes to the methodology. However, there are some differences when it comes to the population since they often are employed at companies, and the types of questions that are specific to software engineering.

Since the methodology relies on sampling, it is typically carried out by first planning and then carrying out the study according to the plan (i.e. a "fixed design" according to Robson [39]). This means that conducting a survey can be divided into a number of sequential steps.

First the research questions are defined based on the need for research, open questions, and interest of the researcher.

Based on the research questions the target populatioon and the audience of the study can be defined. The target population is in most cases large which means that it is impossible to ask questions to each of the members, which means that a sample of the population must be decided as a next step.

Before the sample population can be approached a questionnaire, a sample instrument, must be formulated, which is often done by first defining it and then evaluating it on a subset of the sample population in order to make

sure that it servers its purpose.

When the sample population has answered the questions in the instrument the researcher can analyze the data, and draw conclusions based on the analysis. After that the study with its results and conclusions can be reported.

Each of the mentioned steps are further presented in the subsequent chapters of this report.

Chapter 2

Defining Research Objectives

The first step in the survey process is to *define the research objectives*. These describe the problem or issue of interest, and provide a scope and context for the research questions [25]. They are referred later on in the survey process when decisions have to be made and make up a frame that confines the researchers and stakeholders from steering into irrelevant directions.

When research objectives are specified, certain details will require extra attention. Other than stating the main questions and hypothesizes that the survey aims to investigate on a higher level, the objectives should:

- consider what population of respondents that the survey will target [6, 25]
- explain the motivation behind the survey [6]
- consider other possible directions that fall close to the objectives but are not investigated [27]
- consider what resources that will be needed to accomplish the goals of the survey [27]
- discuss how the information from the end results will be used [25]

2.1 Preparatory work

In correlation to the work of defining the research objectives, a review of related work should be performed [6, 25]. If a previous survey falls in the proximity of the one under development, maybe some parts of the questionnaire could be reused which could increase reliability. It could further help to narrow down the objectives to be even more specific to the goal of the stakeholders.

Alongside the review of related work, discussions should be performed with all stakeholders. Everyone needs to be in agreement on the core questions and what kind of information that can be expected in the end [6]. As with the general size of the scope, unnecessary iterations later on in the survey process will prove expensive and might effect the end result in a negative way.

2.2 Top-down versus Bottom-up process

Formulations should capture the goal of the survey, either as statements of the expected outcome or as questions that break down the problems or issues of interest [27], e.g. with a series of *"What, how and why"* questions. Further fine grained definitions of the research objectives constitute the research questions. This top-down reasoning can be exemplified in a survey by van Heesch et al. [49], about how software architects reason where the goal was *"...to understand the reasoning process that industrial software engineering practitioners follow while they are architecting."*. Van Heesch et al. continues by breaking down the goal into research questions by mapping to existing literature which results in the following research questions.

- *"RQ1: How do software architects scope and prioritize the problem space during architectural analysis?*
- *RQ2: How do software architects propose solutions during architectural synthesis?*
- *RQ3: How do software architects choose among solutions during architectural evaluation?"*

This way of defining of the research questions correlates to Ciolkowski et al. [6] proposition of adopting of the Goal Question Metrics (GQM) method [50]. This is a goal-oriented method used especially in the field of software quality improvements. With a top-down approach, goals are first defined followed by more detailed questions and then in turn metrics, which then are intended to measure to the goals. Transferred to the survey domain, goals corresponds to research objectives, questions to research questions, and metrics to questionnaire.

Kasunic [25] suggests that this also can be done from a bottom-up approach, i.e. that the research questions can both be derived through, and provide support to further help define and narrow down the objectives.

2.3 Different types of objectives

With the GQM method in mind, the researcher needs to be aware of what is to be measured and how this should be reflected in the research objectives and questions. Kitchenham and Pfleeger [27] describe three types of objectives in regards to this:

- Measure a rate or frequency of a certain characteristic among the investigated population. E.g. the frequency of failing projects [16].
- Evaluate the gravity of a certain characteristic among the investigated population. E.g. the average overrun of software projects [33].
- Discover factors which affect a characteristic among the investigated population. E.g. factors that predispose a process improvement activity towards failure or towards success [11].

2.4 Purpose of research

As described by Wohlin et al. [53], the objective of a survey can be either i) descriptive, ii) explanatory, or iii) exploratory. Descriptive, meaning they give the researcher support to make claims or assertions about the population. The questions tend to be more although not exclusively about what, then why. For example: what agile practices are used, in contrast to why are agile practices used? Explanatory surveys help the researcher explain trends, phenomenons or problems observed in the population. It takes a further step and asks the why from the previous example. Exploratory surveys help the researcher to break new ground and discover new insights into an area that is to some degree unknown. It can for example, be used as a pre-study and to render propositions for future research.

As noted, the purpose has implications on the research objectives, also confirmed by Robson [39]. To give another example, now from the perspective of an evaluation of introducing agile practices to a software development organization, the different purposes could render the following research questions:

- Descriptive: What are the developer's view on the new agile practice?
- Explanatory: Has the developers become more efficient after the introduction of the new agile practice?
- Exploratory: How does the introduction of a new agile practice affect the developers?

2.5 Economic considerations

Scope coverage of the research objectives needs consideration from an economic perspective. If too wide, it will open up size and complexity issues later on in the design process of the questionnaire. A consequence could be that wrong topics end up as questions. Even though these might be corrected in a pre-stage, the iterative work needed will consume unnecessary time and resources from the surveys budget.

2.6 Critical perspective

It should be questioned whether survey is the correct methodology to be applied in a specific investigation. Kitchenham and Pfleeger [27] take perspective from the different steps in the survey process. Is it clearly defined what population context that should be of focus when deciding on target population? Is there a reasonable way of sampling the population? Is there a budget that can cover the survey based on what is asked? Is it clear what variables that needs to be measured and how?

Sometimes a survey can only provide a part of the answer. Kasunic [25] refers to the concept of triangulation and proposes that a series of methodologies could be used as complements to obtain the whole picture. This could especially be the case of exploratory surveys as mentioned earlier.

Chapter 3

Identifying Target Audience and Sampling Frame

3.1 Target Audience

When the target population for a survey is identified, it can be considered that its target audience, i.e., who are its intended respondents, is established [25]. A target audience must be established taking into account who can best provide the information needed in order to achieve the research objective. In this sense, it is important to highlight that survey instruments should be written from the perspective of the respondent, not from the perspective of the researcher, applying a vocabulary that must be interpreted by them. Also, a method of surveying (interview, web questionnaire, etc.) must be chosen taking into account which of them could be more accessible for a target audience.

3.2 Characterizing the Target Audience

Kasunic [25] presents a set of basic attributes that can be applied in order to elicit the main attributes on characterizing the target audience in Software Engineering surveys. In addition, we suggest classifying them from the point of view of how the data is collected as *dependent* ("D", typically related with subjects' background) or *independent* ("I", typically including demographical attributes) from the research context:

- size (I)
- jobs and responsibilities (I)
- education level (I)

- gender (I)
- age (I)
- technical abilities (D)
- relevant experience (D)
- perception regarding the survey' domain knowledge (D)

3.3 Population

In statistics, a*population*, consists on the set of accessible elements from the target audience from which samples can be extracted for a specific study [48]. Populations can be divided into sub-populations, each one having well defined and distinct values for a set of properties from each other. In such cases the population can be classified as *heterogeneous*. For instance, considering the population of individuals living in a specific country, an experiment can be performed in order to observe how certain intervention (such as the administration of a drug) can have different effects over different ethnic groups. However, it is important to note that the heterogeneity of a population should be analyzed based on the variables that characterize the context of each research.

Kitchenham et al.,[28] assert that if there is no way to characterize the population from which samples are extracted, no inference over the results can be drawn. In this context, knowledge areas such as medicine offer accessible resources for supporting its researchers on investigating representative populations available for each study. However, in Software Engineering research, there are significant limitations on performing this activity [9]. Thus, alternative sources such as social networks, digital libraries and open databases of software projects have been investigated to reach mitigate such limitations [10].

3.4 Unit of Observation, Unit of Analysis and Search Unit

In questionnaire based surveys, the data is always collected from individuals (*units of observation*) (primary object), necessarily individuals (respondents). However the survey design may demand a higher level of analysis (*unit of analysis*) distinct from the own individual. For instance, Conradi et al. [7] presents a survey in which individuals (unit of observation) that worked in software projects (unit of analysis) using off-the-shelf (OTS) components were surveyed.

Another relevant issue regarding the population is related with how the units can be *filtered* and *retrieved* in a specific *source*, introducing the concept of *search unit* [10]. Ideally, the search unit and the unit of analysis of a survey are represented by the same entity. However, due to the limitation of sources available for sampling, these concepts can be represented by different entities. For instance, in the case of the Conradi et al. survey, to access a large scale of OTS projects (unit of analysis), they randomly selected and sent invitations to representative sets of ICT companies (search unit) from three distinct countries.

3.4.1 Examples

Although SE survey academic papers and technical reports often omit clear descriptions regarding their target audience and population, Table 3.1 presents a set of SE surveys in which it was possible to clearly identify their units of analysis, units of observation and search units. In the case of [7], these three concepts presents distinct values as already mentioned, whereas in the survey presented by [40] these three concepts are represented by the same entity. Alternatively, de Mello et al.[10] and Bettenburg et al.[3] worked with web-based sources, which offered ways of retrieving units of observation from alternative search units in large scale.

3.5 Sampling Frame

In statistics, a *sampling frame* is the source from which a *sample*, i.e. a subset of units from the target audience, can be retrieved [41]. In many practical situations, the establishment of a sampling frame is a matter of choice of the researcher; in others, the sampling frame has a clear critical importance for the interpretation of the study results. Sarndal et al.[41] observe that some appropriate investigations could not be carried out due to the lack of a suitable sampling frames, while other investigations remain inconclusive results due to incomplete sampling frames. For the authors, an *ideal* sampling frame must present the following set of characteristics:

1. All units have a logical, numerical identifier
2. All units can be retrieved and relevant information from them is available
3. The frame is organized in a logical and systematic fashion
4. The frame has additional information regarding its units
5. All elements of the target audience is present in the frame

Study	Target Audience	Unit Of Analysis	Unit of Observation	Search Unit	Source of Sampling
[7]	Professionals from ICT companies that worked in finished software projects using OTS components	finished software projects using OTS components	professionals from ICT companies	ICT companies	Databases of ICT companies from Italy, Norway and Germany
[3]	Experienced bug reporters	developers	developers	bug reports	Mozilla and Eclipse projects database
[40]	Finnish Software Practitioners suited to the survey focus	Finnish Software Practitioners suited to the survey focus	Finnish Software Practitioners suited to the survey focus	Finnish Software Practitioners suited to the survey focus	FIPA (The Finnish Information Processing Association)
[10]	SE professionals experienced with agility in software process	members	members	groups related with agility in software processes	LinkedIn

Table 3.1: Examples of target audiences in SE surveys

6. All elements of the target audience is present only once in the frame
7. No element outside of the target audience are present in the frame
8. The data is 'up-to-date'

Sarndal and Wretman[41] observe that the characteristics 1 and 2 are considered *essential* while the other characteristics are desirable, contributing to the quality of the sampling frame and on reducing the effort involved in the sampling process. The third characteristic, for example, supports simplifying the sampling process, while the characteristics 4 and 5 contribute to simplifying stratifying activities. In fact, when all eight characteristics are met, we can consider that we have obtained an ideal sampling frame. Due to non-conformities with the desirable characteristics described, Kish[26] identifies four basic problems of sampling frames.

1. *Missing elements*, when the sampling frame does not include all units from the target audience;
2. *Foreign elements*, when there are elements in the sampling frame out from the target audience;
3. *Duplicate entries*, when two or more elements from sampling frame represents the same unit;
4. *Groups based on clusters*, i.e. rather than groups based on individuals.

Due to the lack of accessible and controlled sources for establishing representative sampling frames in Software Engineering surveys, it is frequently hard to avoid the aforementioned basic problems and to support the characteristics needed for composing an ideal sampling frame. For instance, De Mello et al. [10] established a sampling frame composed by a small subset of the source (the professional social network *LinkedIn*) following a systematic searching process. Bettenburg et al. [3] study established a sampling frame composed by a small set of open-source projects available.

Chapter 4

Designing Sampling Plan

Sampling is the process in which a *sample*, i.e. a subset of units of observation from a *sampling frame* is selected to be used in a study. Sampling is typically needed when the effort involved on selecting all units (census) from a sampling frame is not feasible, or even when the selection of all sampling frame could bring an collateral effect on the statistical power, introducing a prohibitive hypersensitivity to the sample [22]. Following subsections present the most common sampling designs available in the specialized literature, distributed between *non-probabilistic* and *probabilistic* designs.

4.1 Non-probabilistic Sampling

Non probabilistic sampling is related with all sampling approaches in which randomness could not be observed on selecting the units, i.e., the units from a sampling frame do not have the same probability to be chosen [48]. As main consequence, the extent in which the observed results can be generalized for the whole sampling frame is limited, even when the evidence obtained is statistically significant. Specialized literature presents the following four main non-probabilistic sampling designs:

- Accidental sampling: The only criterion for selecting each unit is the convenience. It is a common design on SE surveys in which, frequently, researchers recruit subjects from their personal connections.

- Quota sampling: The sampling frame is composed by mutually exclusive subsets in which their units do not have the same probability to be chosen since quota sampling does not take into account the size of each subset. For instance, a survey design could establish that 20 companies will be surveyed limiting the survey invitation to only ten employees from each company.

- Judgment sampling: It aims to reduce the bias from the accidental sampling since there are clear reasons for selecting each unit. It includes practices such as the use of experts' opinion on selecting units.
- Snowballing sampling: It extends accidental sampling, typically selecting seeds of subjects (first level) for indicating subjects (second level) to be recruited by the researchers. Alternatively, there is a common approach in which the first level units are allowed for direct recruiting new subjects.

4.2 Probabilistic Sampling

To be considered a probabilistic sampling, all units from a sampling frame must have the same probability to be selected which can be supported through random sampling [48]. As a consequence, it will be feasible to calculate the *confidence level* of the observed results (in which extent these results are reliable) and its *confidence interval* (in which level the results can be extended to all sampling frame).

4.2.1 Simple Random Sampling (SRS)

This is the most common probabilistic sampling design, in which n distinct units are selected from N a sampling frame having all of them N the same probability to be chosen [48]. Thus, considering the establishment of an adequate sampling frame, performing SRS means that all its units can be considered homogeneous from the point of the view of the study scope. It is important to emphasize that, when establishing the sample size for a voluntary survey (frequently in SE) the probability of each recruited subject effectively participate (effective sample size) should be took into account. Such participation may significantly vary depending on the survey plan.

4.2.2 Clustered Sampling

In this sampling design, homogeneous clusters of distinct units can be identified in a population. Then, if the observed clusters present significant similarity levels between then, it can be considered to only select a subset from these clusters for a trial. As a consequence, due to this similarity (identified as function from the units' attributes designed for a survey), a small loss of confidence is expected and, at the same time, significant efforts on recruiting and data collecting can be avoided [48]. Clustered sampling is commonly applied in large scale surveys in which researchers must be *in loco*, (in person) for collecting data [2][38]. As there are still many challenges on characterizing the context in SE studies, including surveys, it may

be considered risky to apply clustered sampling.

4.2.3 Stratified Sampling

In statistics, stratified sampling is considered the best sampling design for large scale populations, allowing distributing all the population units into distinct subpopulations (*strata*, plural of *stratum*). Then, for each *stratum*, SRS must be performed, which allow to observe more specific and reliable results than in a single SRS. It is important to highlight that no units from a sampling frame can be let out from a *stratum* and all *strata* must be mutually exclusive [48], i.e. a single unit can't be found in more than one strata.

4.2.4 Systematic Sampling

Systematic sampling consists of a SRS in which a previously sorted sampling frame composed by N units have a sample of n units selected following a sequence initialized by a randomly selected unit i. Then, the next units are selected through the continuously addition of the interval k, resulted from the integer division between N and n). For instance, if the population size is 200 and the sample size must be 50, k=4. Then, if i=3, the following 10 first units will be included in this sample: 3, 7, 11, 15, 19, 23, 27, 31, 35 and 39.

Chapter 5

Designing Survey Instrument

5.1 Designing Questionnaire

A survey instrument is usually a questionnaire that is very important and requires special considerations. This section presents guidelines to design survey questionnaire and to develop internal and survey questions. Internal questions are open ended questions that are later transformed to survey questions. Internal questions represent main objective or goal of the investigation that is being carried out. The results and conclusions of a survey directly depend on the quality of the used questionnaire. The main strength of survey research lies in the collection of a population's behavioral and attitudinal attributes quantitatively, which allows uniform interpretation of the collected attributes [25]. To achieve sound results and conclusions from the survey research it is important to carefully design the questionnaire of a survey, for this one has to consider the following factors:

1. Following a team based approach
2. Determining what is to be measured
3. Aligning internal and survey questions with the research objectives
4. Selecting questionnaire type
5. Prioritizing internal questions
6. Selecting survey questions type
7. Selecting execution method
8. Questionnaire length
9. Sequence of survey questions

10. Establishing response format
11. Transferring internal questions to survey questions

5.1.1 Following a team based approach

The first task of designing a questionnaire is to form a team of both researchers and domain experts. This team will provide required expertise to design survey questionnaire including both technical and substantive knowledge of the project or topic under investigation [37].

5.1.2 Determining what is to be measured

In the initial phase of designing a survey questionnaire it is good to to know clearly what is going to be measured and because of this the collected data will be easily analyzable. By using survey research one can collect three types of information such as, descriptive, behavioral and attitudinal [37]. Descriptive questions consist of the respondent's age, education, occupation, and experience. The collected information from descriptive questions is also known as demographic or personal information about respondents. Kasunic [25] calls descriptive questions, the questions about attributes and these attributes can be personal or demographic attributes of respondents. Sometimes researchers are interested in collecting behavioral information of respondents. Behavioral questions try to measure difference or change in the respondent's behaviors. Finally, attitudinal questions collect information about respondent's attitudes and opinions. Kasunic [25] distinguishes between questions about attitudes and beliefs. The author defines attitude as personal outlook or orientation acquired through years of experience. On the other hand, Beliefs are the respondent's assessment of what they think is true or false. However, at the same time Kasunic says that the distinction between attitude questions and belief questions is sometimes a grey area. Here, we are not trying to distinguish between attitude questions and belief questions. In these guidelines they will be used in same context as reported in [37].

5.1.3 Aligning internal and survey questions with objectives

After this the survey questionnaire design deals with the development of survey internal questions. The internal questions must be developed in alignment with the main objective of the study that are later formulated to survey questions. While developing internal questions, the researchers or those carrying out the survey must take objectives and target population of the survey into consideration. If the objectives of the survey are not clearly defined then it will be hard to design a useful survey instrument.

While designing a survey instrument one should also consider the target population of the survey because results and conclusions of the survey will be based on the response of the target population.

5.1.4 Questionnaire types

It is also important to consider and select a feasible type of questionnaire that will be used for the underlying investigation. There are mainly two types of questionnaire, self-administrated and interviewer-administrated questionnaire.

Self-administrated questionnaires are administrated by the respondent him/her self and they are executed by ordinary postmail, E-mail or Internet. In this respondents fill in the distributed questionnaire without the help of the investigation team. Therefore, this type of questionnaire requires some additional information at the beginning of the questionnaire i.e. introductory notes, definitions of used concepts, and possible outcomes. Self-administrated questionnaires are less costly and easy to administer. Since respondents fill in the questionnaire by themselves without the presence of any member of research team, it preserves confidentiality and there is no researcher influence while filling out the questionnaire. However, this type usually results in low response rate because respondents are less motivated to respond. To reduce this problem one can carefully design the questionnaire by writing proper introductions with information about the importance of the survey results for respondents.

Web-based questionnaires are becoming more and more popular because they are more time efficient and help to acquire higher response rate than mailed questionnaires. They are easy to set up and then their distribution is also very simple and straightforward by sending the corresponding link to target audience. Data collection by investigators is also easy because it does not require the time consuming data entries. However, the selection of a web-based questionnaire may miss a part of the population that does not have access to internet. While designing a survey questionnaire and selecting questionnaire type one has to consider target population and sample. These are some examples of web-based questionnaire applications; Survey Monkey[1], LimeSurvey[2], QuestionPro[3], and QuestBack[4].

There are two types of interviewer-administrated questionnaire, face to face and telephone interviews. By using them the target population can easily be accessed regardless of their education or expertise (like previously not whole population has access to internet and not everyone can use comput-

[1] http://www.surveymonkey.com
[2] http://www.limesurvey.org
[3] http://www.questionpro.com/web-based-survey-software.html
[4] http://www.questback.com

ers). In this, the interviewer (investigation team member) can help to clarify ambiguous questions and also the response rate is obviously higher than in the web-based or mailed questionnaire. By using this type of questionnaire one must take care of interviewer bias that is introduced by different interviewers' perceptions and interpretations of the answers. Use of interviewer-administered questionnaires in large survey is costly because they require more resources (interviewers), also they are not a suitable choice for the collection of sensitive information.

5.1.5 Prioritizing internal questions

While formulating internal questions from the survey objectives it is important to identify and prioritize important questions [25]. The prioritization of internal questions will help later in keeping a reasonable questionnaire length. As designing a questionnaire and internal questions is a team activity, that can result in a large number of internal questions. Usually it is not easy to prioritize internal questions because it requires great amount of effort and discussions among the investigation team members, but this prioritization can be useful while considering a reasonable questionnaire length.

5.1.6 Survey questions types

There are mainly two types of questions that can be used in a survey questionnaire, open-ended and closed-ended. In Open-ended, respondents are asked to answer without given any answer choices. Respondents answer these questions in their own words subjectively that provides a great deal of openness to respondents while answering. On the contrary, closed-ended questions provide a fixed list response choices or categories and ask respondents to select one or more as their answer. Closed-ended questions are mostly used in a survey questionnaire however both types have their own advantages and disadvantages. Open-ended questions do not impose any restriction on respondents to choose one from already given choices. However, they lead investigators with the qualitative data that is not easy to interpret and analyze for generalizable conclusions. Therefore, if the plan is to have open-ended questions then one must need to analyze available resources. It does not mean that open-ended questions are useless for survey research, they can be very useful in interviewer-administrated surveys [25]. Moreover, there is another commonly used question type in most of the questionnaires that combines both open and close ended questions known as hybrid questions. They are also known as partially closed-ended and they do not impose restriction on respondents to choose one option from already given options because they also provide an option of open response from respondents. If respondents think that the given options are not suitable

then they can write answer in their own words subjectively.

Inclusion of only one type of questions in a survey questionnaire may not be a good choice. for example, if a questionnaire contains only hybrid questions then that can lead to a large amount of qualitative data to analyze, which requires a lot of resources. Here, it is important to find a balance between type of questions while considering research objectives and available resources.

5.1.7 Execution method

The execution of a survey can be carried out by several methods therefore it is important to discuss among the survey team and select a suitable execution method during the designing phase of questionnaire. A survey can be executed by the following methods: ordinary mail, web-based questionnaire, face to face interviews and telephone based interviews. Each of these execution methods have advantages and disadvantage (for details see [37]). Today as survey research has become popular and common, a web-based execution method is the first choice in the most cases. In a software engineering context it is common to execute surveys through a web-based questionnaire because mostly it involves software engineers or programmers that, in most cases, have access to internet and can use computers to fill web-based questionnaires. In some cases researchers evaluate the usability of a system or software by targeting users of that system or software. This means, it is likely that most of the users of the system under evaluation or investigation are able to use computers and have access to internet. Sometimes in an organization employees are asked to fill in paper based questionnaires, this method gives ease of collection of data if it carried out in an organization with limited number of employees. If one intends to use a paper-based questionnaire to target a whole population then that will involve mail cost and also require huge resources to manually enter that collected information in a digital form to analyze it. This section emphasizes on the careful selection of execution method by keeping in mind the available resources and by clearly knowing what is intended to measure. The web and interviewer based execution are explained in the previous subsection 5.1.4.

5.1.8 Questionnaire length

The questionnaire length is very crucial because it directly affects the survey response rate [54]. It is common practice that investigators distribute lengthy survey questionnaires among participants to investigate unclear and not well defined research objectives. If the research objective is not clear and not well defined or its scope is too large then it is likely that the questionnaire will be very lengthy. Some studies have presented the factors that

effect the response rate in mail and web based surveys [13, 54, 19]. For example, it was observed that every additional question can reduce response rate by 0.5% and every additional page by 5% [19]. If a questionnaire is longer than 4 pages then it has a significant effect on the response rate [54]. This is also valid for a web based questionnaire. As stated earlier, the prioritization of internal questions help in limiting the length of a survey questionnaire. If a questionnaire is becoming very long then one can remove some questions from the performed prioritization to keep it in a reasonable size. It is also important to show the total number of pages of a survey questionnaire on its first page. The respondents should clearly know about the length of questionnaire, sometimes they become bore or tired by filling out few pages and still they do not know how many pages are left.

5.1.9 Sequence of survey questions

The arrangement and organization of survey questions in a survey instrument are crucial and require special considerations. The order of survey questions in a questionnaire can greatly effect the results of a study, specifically it can effect the response rate. If the survey questions are poorly ordered and organized then they can confuse respondents, and demotivate them to fill questionnaire that consequently can affect the entire research effort and yield to poor results [37].

Kasunic [25] presented three categories of questions that can be included in a survey questionnaire,

- Demographic questions
- Substantive questions that are the main questions of a survey
- Filter questions

Moreover, one more question category is mentioned by [37],

- Sensitive questions

Demographic questions are also known as introductory questions and the main objective of these questions is to motivate respondents to continue with questionnaire without confusing and demotivating them [37]. Therefore, they should come early in a questionnaire. These are about respondents characteristics, for example, job, experience, gender etc.

Substantive questions address the main objective of survey, they collect information regarding the topic that is being investigated. It is important to have all substantive questions together, which can gain full concentration of respondents without any distraction. If the objective of the study is broad

and covers more than one subject in particular then it is good to introduce separation by different categories of questions. Each category should have its own heading and a short introductory note about that section.

Filter questions require respondents qualification to answer subsequent questions, they do not apply to all respondents. They are also known as screening questions. For example, in a software engineering context if a respondent has more than three years of experience of development than he/she has to answer subsequent question.

Sensitive questions collect sensitive information, such as religious affiliation, ethnicity, sexual practices, income and opinion about highly controversial ethical, political and moral issues [37]. In a software engineering context sensitive questions can be about respondents salary, and opinions about organization or management. If these questions are not placed at the right place in a questionnaire then they can negatively effect the results. Respondents can react negatively by providing sensitive information and decide not to participate in survey, resulting in low response rate. By considering this threat it is important to place these questions at the end of questionnaire. In a software engineering context, these questions are not common but it is still good to keep them under consideration while designing a questionnaire.

5.1.10 Response format

In survey research data is collected in form of different variables that represent specific characteristics of the respondents, such as, age, job, experience, skills etc. There are mainly three ways in which these collected variables can be measured,

- Nominal or dichotomous response scale
- Ordinal or Likert response scale
- Interval response scale

In the nominal response scale, data can be placed into categories, e.g., yes/no, agree/disagree and can only be measured regarding frequency. For example,

Q. Have you used Eclipse?

Yes——— No——— Do not know———.

The 'do not know' option is presented for a respondent to keep motivated and interested. If a respondent does not want to answer a particular question or does not know about it then she can choose this option.

In the ordinal response scale, data placed in categories can be ranked according to their order but without indicating the magnitude of differences between them [37]. This response scale determines the intensity of a belief or opinion and frequency of a behavior [25]. For example,

Q. How difficult was Eclipse to use?

Very Easy——— Easy——— Fair——— Difficult——— Very Difficult———.

In the interval response scale, constant units of measurement are used to indicate exact values of each category response. Income, age, and experience, etc. can be measured using interval response scale. For example,

Q. How much experience do you have of using Eclipse?

1. ———Less than 1 year
2. ———1 year to 3 years
3. ———3 years to 5 years
4. ———5 years to 8 years
5. ———More than 8 years

5.2 Designing survey questions

The wording of survey questions is critically important to achieve useful results and conclusions from carried out survey. In general, there are three conditions that must be fulfilled to get appropriate response of survey questions from their respondents,

- The questions must be understandable by the target population
- Making sure that respondents have sufficient knowledge required to answer survey questions
- Are participants motivated and willing to participate in survey?

If afore-mentioned conditions are not met then one of the following problems can arise. First, the collected information from the carried out survey could be inaccurate. Second, there could be many responses with"do not know" option, resulting in incomplete answers of questionnaire. Finally, participants could refuse to participate in the survey by lack of understanding of questions or by not having the required knowledge to answer questions. To avoid these problems it is important to pay special attention while writing survey questions. For this, it is recommended to take care of the following question wording problems, [25, 37, 27]

- **Using appropriate and simple language**: It is important to use simple and appropriate wording for survey questions. Here, appropriate means keeping the target population in consideration while writing survey questions that all respondents can understand them. Always define any likely ambiguous terms used in survey questions.
- **Avoiding technical terms**: It is likely that a survey in a software engineering context can have some technical terms that are not well known by all respondents (e.g., software engineers, developers, testers, etc.). In such case it is important to either avoid those technical terms or define them in the introduction section of a survey.
- **Keeping questions short**: It is important to have short questions that ask about only one concept.
- **Avoiding vague sentences**: It is important to avoid too vague sentences while writing survey questions.
- **Avoiding biased questions**: Biased questions suggest likely answers or responses in their sentences. It is important to avoid biased questions, which can be done by carefully phrasing the questions that do not suggest likely answers or responses.
- **Avoiding sensitive questions**: Special care must be taken while writing survey questions, they should not have sensitive questions (too personal, i.e., sex, income, etc.) that can lead to low response rate. In a software engineering context, the sensitive questions can be about respondents income, opinion about organization or management, etc. If for some reason one wants to ask a few sensitive questions in a survey then they should come at the end of the questionnaire.
- **Avoiding too demanding questions**: It is important to avoid too demanding (that require too much effort at the respondents end) questions while writing survey questions.
- **Avoiding double-barreled questions**: It is important to avoid asking two questions in one question, which is known as double barreled or composed questions. If a researcher asks these questions then respondent most likely can not answer both questions asked in one question, which leads to difficulties in analyzing collected data.
- **Avoiding double negatives**: Double negative questions are confusing questions and they must be avoided while writing survey questions.
- **Avoid asking about past events**: It is important to avoid questions about events that have occurred long time ago. In this case it is likely that respondents do not remember exact information, which is being asked in survey question.

Chapter 6

Survey Instrument Evaluation

This section presents guidelines for designing and conducting pretest or pilot surveys to evaluate survey instrumentation. Designing a survey instrument is an iterative activity that involves an evaluation though carrying out a pilot run to introduce improvements in the survey instrument. This way, survey instrument can be improved iteratively by getting feedback from a pilot run of a survey.

The output of the previous step, designing instrument, is a draft questionnaire that requires a pretest or a preliminary evaluation. Pre-test evaluation is also used to assess the reliability (reproducing the similar survey data while administering survey again) and validity (is survey measuring what is required to measure?) of instrumentation [27] (For more details see chapter 8). The survey evaluation mainly assesses the following critical factors [37, 27],

- Questionnaire clarity and understandability
- Likely response rate and its acceptability
- Effectiveness of execution method and other survey supporting documents
- Reliability and validity of survey instrument
- Data collection and its analysis

6.1 Survey instrument evaluation methods

A survey instrument can be evaluated by using one of the following evaluation methods [36],

- Expert reviews
- Focus groups
- Pilot surveys
- Cognitive interviews
- Experiments

6.1.1 Expert reviews

The first type of evaluation of survey instrumentation is an expert review. There are two types of expert reviews, survey design expert reviews and subject matter expert reviews.

The survey and questionnaire design experts evaluate the survey instrumentation and make sure that it is designed according to the best practices in survey research.

The subject matter experts evaluate the survey instrumentation and help to find the correct survey questionnaire flow. They check the wording used in the questionnaire and determine is that technically correct or not. They help to align the survey instrumentation with the survey main objectives or goals. Moreover, they determine the understanding of survey questions that they will be understood in the same way by all respondents and this understanding matches with the survey designers intention.

6.1.2 Focus groups

The most common types of evaluation used for a survey instrumentation are focus groups and pilot studies. A focus group is a qualitative research component in which a group of people are asked about their perceptions, opinions, beliefs, and attitudes towards an under investigation topic or project. It usually consists of people representing survey investigators and participants. It evaluates instrumentation and help to identify missing or unnecessary questions and ambiguities. In focus groups, questions are asked in an interactive group manners, often face to face, where group members are free to talk with other group members. In focus groups, the investigator or moderator gathers a group of people (usually 7 to 8) and asks them survey questions. It allows focus group participants to provide longer answers and discuss a topic with others. A focus group can also be used and helpful in gathering information during designing a survey questionnaire to see what topics are important to a sample of the population, how people understand a topic area and how people interpret questions. It provides a qualitative understanding of the topic under investigation that is being quantified in a survey research.

6.1.3 Pretest/Pilot surveys

Pilot studies are carried out by using the same material and procedures (designed for the final survey) including survey questionnaire, execution method, questionnaire format, motivational documents, etc. but with a small number of participants from the target population. Pilot studies evaluate survey instrumentation like focus groups but additionally they also evaluate the response rate and follow-up procedures. Pilot surveys are usually conducted well in advance before carrying out the final survey so that more substantial changes to the questionnaire or procedures can be made. Carrying out a pilot study can be time consuming and requires more resources than a focus group meeting. However, pilot study provides a realistic opportunity of improving survey instrumentation before carrying out the whole designed survey. Carrying out a pilot study in software engineering context is as important as it is in other fields [30].

6.1.4 Cognitive interviews

A cognitive interview is not like an ordinary interview. Ordinary interviews produce codeable responses to the interview questions. On the other hand, cognitive interviews produce a view of the processes evoked by the survey questions [36]. Cognitive interviews help to identify and analyze sources of response errors in survey questions by focusing on the cognitive processes, which respondents use to answer the questions [52, 18]. The main objective of these interviews is to focus on the survey questions rather than on the respondent. This way, these interviews are used to evaluate the quality and accuracy of the survey instrumentation [52]. To the best of our knowledge no one has used cognitive interviews for the survey research in a software engineering context.

6.1.5 Experiments

Experiments can be conducted to evaluate the survey instrumentation. The expert reviews, focus groups and pretests lead to revisions in the survey instrumentation. After carrying out revisions, it becomes difficult to determine which version of the survey instrument is better that should be used for the final designed survey. Experiments can compare different versions of survey instrumentations and help survey designers to select the best version. Experiments can be used to compare both the single item of a questionnaire and the entire questionnaire [34]. This way, by conducting an experiment, one can evaluate different versions of the survey instrument and determine which version to use in the future.

Chapter 7

Analyzing Survey Data

As already mentioned in Chapter 5, each survey question can be classified as *closed*, *open-ended* or a mix of them. While the analysis of the answers for closed questions are better supported through *quantitative methods*, open-ended questions are better suited to be analyzed through *content analysis*. Both cases are addressed in the following sections.

7.1 Open-ended questions

Open-ended questions are designed in such a way to encourage the explanation of the answers and the reactions to the question through a sentence, a paragraph, or even a page or more, allowing researchers to better access the respondents' true feelings on an issue. Following subsections introduce the *content analysis* approach for analyzing survey data extracted from open-ended questions.

7.1.1 Content Analysis

Content analysis is a method for finding valid and replicable evidence from a *textual data* to their *context*, describing a family of analytical approaches ranging from intuitive (interpretative) analysis to systematic (strict textual) analysis.

Aiming at attaining a condensed and broad description of the phenomenon, delivering categories and concepts through a *coding process*, content analysis can deliver knowledge, new insights, representation of facts, and practical guides to action.

Usually, the concepts and categories support the designing of models, conceptual systems and conceptual maps. Although the criticisms that content analysis does not support a detailed statistical analysis and that content analysis is not sufficiently qualitative in nature, currently is recognized the

classification of two main types of such method, introduced in the following subsections: *qualitative content analysis* and *quantitative content analysis* [12][20].

In fact, many other methods for analyzing open-ended questions can be applied, but they are better suitable for the context of researches typically used in studies based in unstructured/ semi structured interviews. Examples of such methods are Phenomenology [5], Grounded Theory [14], discourse analysis [42] and metaphor [43].

Quantitative Content Analysis

Quantitative Content Analysis (QtCA) is a *deductive* research approach with *positivist* orientation. It means that the phenomena are observed and its reasoning are performed following one or more statements (premises) to reach a logically certain conclusion. QtCA was the first approach applied in research for content analysis [20]. It aims to make replicable and valid inferences from texts to the context of their use [51].

Since QtCA is based on previous research, allowing the formulation of hypotheses about relationships among variables, QtCA coding scheme must be objective, being determined *a priori*. In QtCA coding scheme, relevant and valid categories must be established to support hypotheses testing.

A good coding scheme has exhaustive and mutually exclusive categories, i.e., all relevant aspects of the construct are represented and they are clearly distinct. It is expected that a coding scheme can be measured in the four scales of measurement (nominal, ordinal, interval, and ratio). Also, it is expected to have clear definitions and easy-to-follow instructions. It is important to highlight that if it is observed that the coding scheme must be modified during the coding, it must be re-applied to all the data already coded [51].

In the analysis, many statistical approaches or techniques can be chosen, taking in consideration not only the questions addressed but also the nature of the data, such as: tabulations, associations, correlations, multivariate techniques (hierarchical clustering analysis, multiple regression and others)and semantic nodes [51]. In their work, de Mello et al. [8] applied a coding schema aiming at categorizing the five main skills in Software Engineering reported by each subject in a survey on Agility in Software Processes. This approach was helpful to distinguish the similarities of skills between subjects from distinct set of respondents (*strata*). For supporting this analysis, hierarchical clustering analysis was applied.

Qualitative Content Analysis

Qualitative content analysis (QlCA) is an *inductive* research approach, i.e., from specific observations it is possible to derive broader generalizations and theories. QlCA is based on *naturalist/humanist* research orientation in which research questions guide data gathering and analysis, but potential themes and other issues may arise through data reading. QlCA aims to capture the *meanings*, *emphasis* and *themes* of messages, also understanding the organizations and process on how they are presented. It searches for multiple interpretations by considering the diversity represented by aspects such as ideological positions, critiques or the diverse use of the texts examined [51]. The coding Schema for QlCA follows a subjective approach, in which no previous categories are established. In order to draw categories, the research must look for diversity of ideas, alternatives, perspectives, oppositional writings and different uses of the data sources [51].

In the analysis, QlCA is deeply oriented on grounding in the data, "supporting their interpretations by weaving quotes from the analyzed texts and literature into their arguments and conclusions, by constructing parallelisms, by engaging in triangulations, and by elaborating on any metaphors they can find" (Krippendorff, 2012) [29].

7.2 Closed-ended questions

Closed-ended questions provide a fixed list of response choices or categories and ask respondents to select one or more as their answer. These type of questions are most commonly used type in the survey research [25]. They collect quantitative data and there are mainly three ways in which this collected data can be measured, nominal or dichotomous response scale, ordinal or Likert response scale, and interval response scale. The questions produce quantitative data, which can be analyzed by using methods presented in the following subsections.

7.2.1 Pre analysis considerations

The methods that will be used to analyse any quantitative data produced by a survey is implicitly decided in the beginning of the survey process, e.g. by sampling method and type of research objectives.

The choice of a non random sampling method implicates that no results can be considered representable for the population, from a statistical point of view. Although many of the statistical analysis methods can still be used, any findings will be limited to the respondents obtained through the sample frame.

The type of research objectives implicates the approach in how the analysis should focus [39]. If they are descriptive or exploratory, a more general analysis can be done by using the necessary techniques to present frequencies and possible relationships. If the objectives are more towards the explanatory, then there typically would be a certain set of relationships of pre-defined interest that would be investigated more thoroughly.

To efficiently analyse the data a software with statistical analysis functionality is highly recommended. The option of manual analysis is of course always available but can easily produce errors and will prove inconvenient with even small proportions of data. In many cases, a general calculus desktop software as Microsoft Excel, MiniTab or JMP tools can be used to great success, but for the more complex types of analysis specialized software as Matlab, R, Octave or IBM SPSS is recommended.

When designing the questionnaire it is important to understand how and in which format the data will be collected. Some software may allow an automatic collection of the data if an appropriate interface is available. Others may allow the import of the data structured in certain file formats, e.g. comma separated files. Manual input should be regarded as a last resort as this may introduce errors. If manual input is used, it is recommended that two separate individuals input the data and that it then is cross verified to ensure correctness. A general rule is to have as few steps between the source of the survey data in its original raw format and the final storage and format from where it will be analysed as possible. Reason is that data may be lost, corrupted or ill modified e.g. due to reformatting or data transmission.

7.2.2 Data validation

Validation of the data should be done in an early stage to find any errors as these can later affect the analysis in an harmful and unnecessary way. Steps could include manual proof reading of the data but also by using simpler functionality available in the software. This could be to reorder answers to find erroneous answers or error codes, unexpected number of blanked or extreme answers. Creating frequency tables and graphical representations as histograms or box plots could also help unravel ill-regularities, outliers and extreme values.

This exploratory approach is also a useful way for the researcher to familiarize with the data and get a sense for what it contains. Independent of the nature of the research objectives, it will helpful for the further analysis to have a basic understanding of the data available, how it is formatted, ranges of answers and other general characteristics.

7.2.3 General statistical analysis methods

To help get a better comprehension about what the data represents and interpret what it actually means, measurements from the notion of descriptive statistics may be used. Mean average, median and mode values give a sense for the frequency or answer which could be best representable for the respondents, also known as measures of central tendency. Standard deviation, variance, range and inter-quartile range help to interpret the variability and spread of the answers in the distributions. Depending on the scale used, different values should be used, see table 7.1.

Table 7.1: Summary chart of descriptive statistics that can be used in correlation to scale. The measurments are cumulative downwards. Adopted after Wohlin et al. [53]

Scale type	Measure of central tendency	Measure of dispersion	Measure of dependency
Nominal	Mode	Frequency	
Ordinal	Median	Interval of variation	Spearman corr. coeff, Kendall corr. coeff.
Interval	Mean	Standard deviation, variance, range	Pearson corr. coeff.
Ratio	Geometric mean	Coefficient of variation	

For more thorough explanations on these concepts and related terms, please refer to books on statistical theory, e.g. Siegel and Castellan [45]..

Sometimes it may be necessary to re-scale the data for it to be more understandable or to be able to compare with other results [39]. Subtracting the mean or median from all results gives a better relative view of how far off all values are. Multiplying with a constant for example could be used to transform monetary values between currencies. When the distribution is wide spread or otherwise asymmetrical it could be useful to take the logarithm or power to make it more interpretable. Another tool, referred to as standardization or normalization, involves subtracting the mean from a score and then dividing with the standard deviation. This method is commonly used in statistics and makes it easier to compare distributions with each other.

Correlations can sometimes be identified in the distributions and matched to different types of equations describing a graph. One of the most com-

monly looked for is a linear correlation where the data can be fitted along a linear graph. The fit is seldom perfect why there are different correlation coefficients (e.g. Pearson's r, Spearman, Kendall's Tau A-C) which indicate how good the graph fits the data. As with the measures of centrality and dispersion, it varies when you may use these measures of dependency, see table 7.1. When the distribution is normalized and the scale is either interval or ratio, Pearson's r is used. In the case where the data is not normally distributed, Spearman or Kendall's Tau is to be used. For more information on this topic, refer to e.g. Siegel and Castellan [45].

Often there is an interest to investigate patterns between how respondents have answered in different questions. To find these patterns and correlations there are several tools available in the field of multivariate statistics. Contingency tables, also known as cross tabulation tables, is a matrix which presents the frequency distributions for two variables and how they correlate [39]. This technique can be used for more variables also, but will quickly become cumbersome to present. Others include principal component analysis (PCA), cluster analysis and discriminant analysis. For more information on this topic, refer to e.g. Manly [32] and Kachigan [23, 24].

7.2.4 Hypothesis testing

Given that a relationship can be identified, a null hypothesis can be formulated. This can then be tested in order to determine if it can be rejected or not, and to what level of significance. Choice of test method depends on the type of distribution and its scale.

The tests are generally divided between parametric and non-parametric tests. Parametric tests need the data to fit on a specific distribution dependent on the test. In non-parametric tests, no assumptions are made about the population or the distribution of the parameters. In regards to the null hypothesis, this implicates that for parametric tests it based on parameters of the populations distribution. In non-parametric the null hypothesis is free from parameters.

In regards to scale, the non-parametric tests only exists for nominal and ordinal data, whilst no parametric tests exists for nominal data.

Parametric tests include e.g. t-test, F-test, z-test and ANOVA, whilst non-parametric include e.g. mann-Whitney, Wilcoxon, Kruskal-Wallis and Chi-2. For more information on the topic, refer to e.g. Siegel and Castellan [45], and Good [15].

7.2.5 Presentation and visualization

Frequencies of the answers can be presented in multiple ways. Depending on audience and type of data a consideration has to be made in either presenting

the responses in a table or a graphical format. It is also important to consider the risk of loosing data when choosing a representation. For the more general illustrations graphical representation is usual to prefer. Pie and bar charts are good options when the variables are of discrete character. For continuous variables, histograms and line graphs are to prefer.

As with frequencies, graphical representation of the descriptive statistics helps the interpretation and analysis process. Error bar charts can be used to represent the mean average in relation to the values which lies inside the standard deviation or a certain confidence interval. Box plot is an alternative based on the median value and different quartiles.

Scattergram is a good option to graphically illustrate the data when the variables are of ordinal character. Each axis will represent the frequency for one of the variables. From the plotted distributions possible patterns and (dis)alignments can be identified. As earlier described, it may prove useful to re-scale the data for the plot to make more sense. By the use of different markings, stratas can be graphically separated so their responses can be easier grouped together.

7.2.6 Analyzing Likert items and scale questions

Likert type questions are used to measure respondents attitudes to a particular question or statement, and they produce ordinal data, i.e., one response is higher than another not the distance between the points [21]. This type of question format is common in empirical software engineering surveys why a special subsection is dedicated to this topic. Here are some examples of Likert items,

Q. How difficult was Eclipse to use?

Very Easy——— Easy——— Fair——— Difficult——— Very Difficult———.

OR

Q. Eclipse was easy to use?

Strongly agree—— Agree—— Neutral—— Disagree—— Strongly disagree——.

As Likert type questions produce ordinal data, parametric analysis techniques can not be used. For example, they can be analyzed by using following analysis techniques,

- Mann whitney test
- Kruskal wallis test

A distinction needs to be made between Likert items and a Likert scale. The notion of items refers to a single question with Likert style answer alternatives. The second concept refers to a series of Likert items which aim to measure the same concept. There are separate ways of measuring these two as mentioned in [21, 4].

Single Likert items should be analysed with the median or mode as a measure of central tendency, frequencies as a measure for variability, and Chi-square, Kendall's Tau B or C for measures of association. Examples of single likert items are shown above.

On the other hand, Likert scales should instead use the mean as a measure of central tendency, standard deviation as a measure of variability, and Pearson's r, ANOVA, t-test or regression for other statistical measures. The following questions are used in a previous study [47] that present an example of Likert scale that can be analyzed together to know the opinion of respondents about a risk analysis method. To read more about analyzing Likert scale see for example [21].

Q1. How difficult was the risk analysis method to use?

Very Easy——— Easy——— Fair——— Difficult——— Very Difficult———.

Q2. How difficult were the instructions to perform risk analysis?

Very Easy——— Easy——— Fair——— Difficult——— Very Difficult———.

Q3. You are confident that you have found all the relevant risks.

Strongly agree—— Agree—— Neutral—— Disagree—— Strongly disagree——.

Q3. You will recommend this risk analysis method to others.

Strongly agree—— Agree—— Neutral—— Disagree—— Strongly disagree——.

Chapter 8

Drawing Conclusions

After the respondent data has been analyzed and results are presented, one has to contemplate whether any conclusions can be drawn. It may very well be that the researcher has confidence in the results but this subjectivity is not enough. The complete survey process and its results need to be evaluated and reviewed from a critical viewpoint. The two notions of validity and reliability (e.g. [27]) are central to understanding the thoroughness and trustworthiness of the survey.

Validity in its broadness refers to whether the questionnaire measured what it was supposed to. Reliability regards if the results can be generalizable, i.e. similar distributions of answers would be obtained if the questionnaire was executed in another time, place or sample of the same population. Threats to the validity and reliability can be many and appear in any part of the survey process, e.g.:

- Conflicting research objectives and questions resulting in wrong questions being asked.
- An ill-defined target population resulting in wrong people being asked.
- An improper or badly executed sampling method resulting in a non representative sample.
- Ambiguous and over technical questions confusing the respondents.
- No pilot study executed resulting in missed errors.
- A non user friendly interface to the survey resulting in questions being jumped.
- Data gathering process not pre verified and tested leading to missed data.

- Response rate too low for any valid conclusions to be made.
- Improper analysis methods to the types of questions asked.

8.1 Validity

Each question in the questionnaire should be designed to measure a specific purpose. If it cannot be safely concluded that it measures what it is supposed to then it will be hard to draw any conclusions at all. Here, different types of validity are discussed (topics adopted from Kitchenham and Pfleeger [27]).

- Face validity involves a lightweight review of the questionnaire by randomly chosen respondents [39, 27]. This is an extremely subjective method but can be used to give initial feedback. Should not be used as the single method to support the arguments of validity.
- Content validity, as defined by Kitchenham and Pfleeger [27], is achieved by having a group of reviewers go through and evaluate the questionnaire. The group should include subject matter experts as well as example respondents from the target population.
- Criterion validity refers to how the questionnaire can separate between respondents that belongs to different groups [27]. To measure this attribute an existing classification and mapping of the different groups in the target population must be in place.
- Construct validity is how well the question actually measures the construct it was intended to by the designer [39, 27]. By executing pilot studies and taking concern to content validity with focus groups as earlier mentioned, this can help to ensure the construct validity. Kitchenham and Pfleeger [27] furthermore mention the two variants of construct validity; Convergent and Divergent. The first variant explains to what degree multiple questions intended to measure the same concept actually do. The second type inversely explains to what degree the questions do not correlate, although still intended to measure the same concept.

8.2 Reliability

If conclusions are to be drawn on the whole population, not just on the sample, the reliability needs to be proven and established. Even when just analysing the sample there needs to be a certain level of reliability in order to make any final claims. As with validity, the notion of reliability is broad in its definition. In literature it may also be termed as external validity

[6] and generalizability [39]. Here different ways of measuring reliability is discussed (topics adopted from Kitchenham and Pfleeger [27]).

- Test-retest reliability is based on that the same subject responds to the same survey two times, and it is measured whether the subject gives the same answers each time. Kitchenham and Pfleeger [27] states that if the correlation between both of the answers is greater than 0.7 the test-retest reliability can be considered good. Many factors however effect the actual 'truth' of this test. If given to short of a period between the tests occurrences, the respondent might remember what answer was given last time. On the other hand, if given too long of a period in between, external or personal circumstances may have persuaded the respondent to answer differently the second time [6].

- An alternate form of reliability concerns testing whether the phrasing or reorder of questions has any effect on the answers by a respondent [27]. This method addresses some of the issues with the respondents bias [39] as described in the previous point, primarily in regards to memory. The rephrasing is however a delicate manner is it can trigger other associations or interpretations by the respondent.

- Inter-observer reliability is directed to the observer bias [39] that may be introduced when an observer would involve subjectivity during the conduction of the survey. This regards situations when the survey is not self administered. Another occasion could be during the analysis phase when interpreting and decoding open ended questions. A way to address this validity issue is to have two or more observers involved in the interview and analysis process.

8.3 Risk management

In the beginning of the survey process as many possible threats should be identified and documented [6]. This process (e.g. a focus group) should involve experienced practitioners of the survey methodology as well as subject matter experts and example respondents from the identified population. Those threats that can be managed with redesigning the survey should be looked after. Threats that cannot be managed from an initial point and risk to occur later on in the process should be highlighted and have a cause of action assigned, explaining how to manage these to best extent.

Eliminating a threat to validity or reliability may always have its drawbacks. Better, is of course to use multiple methods if possible in order to rule out as much as possible of any potential effect the threat may have on the results [39]. The cost of resources should however be considered in

correlation to the value added. A limit of when the reliability and validity can be considered satisfactory should be determined by the stakeholders and researchers in the beginning of the survey process. It should also be continuously discussed and adapted as the survey progresses, as different events may take place which can affect what is considered acceptable.

Chapter 9

Documenting and Reporting

From the start, an awareness needs to be in place about what is expected in the end of the survey process. Documentation is one way to start producing deliverable in an early stage as well as a way to increase the general validity of the survey. Then, knowing how to format the results you will can meat expectations more easily and make a higher impact. In this section we both parts will be discussed and presented.

9.1 Documentation

Documenting the survey process helps to increase the quality and acceptance of the survey. It can be viewed by researchers and stakeholders to keep track of where they are, and what step is next. Especially when writing the report, the documentation will come to good use as details can easily be forgotten during the process e.g. if long time passes between the different steps.

The documentation starts in the first step of the survey process, with the specification of the research objectives. The document should then evolve with each step and update iteratively. The content should include at least:

- Research objectives and research questions
- Organization of those conducting the survey
- Description of target population
- Description and design of sampling method
- Any quality assurance steps, e.g. pilot studies or focus groups
- Survey questions to be asked
- Method for distribution

- Method for data collection and analysis
- Schedule of when the different activities should occur, and the different milestones
- Expected results and deliverables

Kitchenham and Pfleeger [27] call this document a questionnaire specification whilst Kasunic [25] refers to its as a survey plan. As the survey respondent data gets analysed, this should also be included in the documentation.

The documentation should also be seen as an agreement between the stakeholders so there are no misunderstandings to what can be expected and when. Although disputes may still arise, a written description of the survey process and its content will ease managing any conflict and avoid unnecessary interruptions in the survey process.

9.2 Reporting

When analysis is done and conclusions are drawn it is time to report the findings. Depending on the venue and audience this packaging of results can be done in many ways. Before starting on the report, the expectations and interests of the target audience should be identified. Kasunic [25] propose conducting an audience analysis. Important aspects to take into consideration in regards to reporting include:

- Form and complexity of language
- Level of detail and abstraction
- What to include or not

The report structure is also dependent on the target audience. In general the following topics should be considered for inclusion (e.g. [6]):

- Abstract and/or Executive summary - The abstract is a short summary of the survey process from start to finish, and should reflect your report. It covers main purpose and scope, objectives, methodology and key conclusions. The abstract should be a quick read for anyone interested and limited to a paragraph with about 150 words. If aimed to management or executives it should be referred to as the Executive summary and can be more thorough.

- Research objectives and problem statement - This is the introduction to your survey report. Together with your research objectives and problem statements should also be a background and scope section which frame the survey to a context and its purpose.
- Methodology and survey process - For the survey to gain creditability, it needs to be thoroughly reported how it was conducted. This is because people need to know under what conditions the results presented were obtained, in order to judge whether they are valid or interesting for their need or interest. This part is more deeply discussed in the next section.
- Results from data analysis - Here you share the results from your analysis of the survey data. The way in which this is done depends on the audience and purpose of the report. Some may require simplified graphical charts whilst others expect percentages and statistical parameters in tables.
- Discussion of results, and threats to validity and reliability - After presenting your results, they need to be discussed and framed to clarify their implications. How did the survey fulfill its research questions? As with previous sections, it is very important that this is done from the point of view of the target audience and their context. The discussion should also frame the whole process from a critical stand point and contrast the validity and reliability of the different steps performed.
- Conclusions and acknowledgements - In the end you should conclude on the hard facts and implications which can be drawn from the survey. This should be in a condensed and easy to read form as the discussion parts has already been performed.
- Eventual appendices - Finally, you should attach any material which could be of interest to the stakeholder and strengthen the creditability of the survey. Questionnaire could be one example.

9.3 Credibility

For the report and its results to be trusted and gain acceptance by its readers, the methodology and description of the conducted survey process needs to be detailed and informative. If the readers do not trust the results, it will get no spread or impact which could cause all effort to be for nothing. Therefore, the notion of credibility is to be seen as equally important as validity and reliability [39], described in earlier chapter. Broken down, creditability can viewed from the terms thoroughness and trustworthiness. Thoroughness being the description of the different steps (e.g. defining target population,

design of sampling plan) and trustworthiness being the description of validity and reliability. Stavru [46] examined the thoroughness and trustworthiness of a series of industrial surveys on agile method usage. He constructed a framework around these two topics as none could be identified in earlier literature.

The topic of thoroughness was based on established survey literature in software engineering [6, 25, 27] as well as in more general fields [31, 35, 44]. The final list contains 21 criteria, see table 9.1, which can be considered relevant. These are weighted between one (least important) to five (most important). When using this list to measure the thoroughness of a survey it is decided whether the survey fulfills the criteria or not. The weights of the fulfilled criteria are then summed and normalized, resulting in a decimal number between zero and one. The closer to one, the more thorough the survey can be considered. Stavru established a limit of 0.62, surveys with a score higher than this would be further analyzed for trustworthiness.

The attributes of trustworthiness was based on those defined by Guba [17]: truth value, applicability, consistency and neutrality. These are mapped by Stavru to the same 21 criteria he had earlier identified. When assessing the trustworthiness, all criteria mapped to each of the attributes, which can be considered to have a positive effect are summed up and normalized. As with thoroughness, this results in a decimal number between zero and one. The closer to one, the more trustworthy the survey can be considered. For more information on the framework, please refer to [46].

Table 9.1: Criteria for thoroughness. Adopted from Stavru [46]

Criteria for thoroughness	Description	Weight
Survey definition		
Objectives	The study defines explicitly its objectives.	1
Sponsorship	The study clarifies its sponsorship (or the organizations who are funding the survey) and their interests in conducting the survey.	1
Survey method	The study specifies and thoroughly describes how the survey was conducted (in terms of its phases, settings and context, etc.).	4
Survey design		
Conceptual model	The study specifies and thoroughly describes its conceptual model (in terms of objects that are investigated, variables and expected relationships between them, etc.)	4
Target population	The study specifies and thoroughly describes its target population (in terms of unit of analysis, reporting unit, exclusion/inclusion criteria, sources, etc.).	4
Sampling frame	The study specifies and thoroughly describes its sampling frame the actual set of units from the target population from which a sample would be drawn (or lists all those within the target population who can be sampled). For example the target population might be defined as all organizations which are developing, maintaining or integration software products and services in a given region. However the sampling frame might be restricted to these organizations which have an official (as the survey would be mediated by email). The information on the sampling frame should include at least the number of units to be sampled.	5
Sampling method	The study specifies and thoroughly describes its sampling method (e.g. non-probabilistic sampling methods, probabilistic sampling methods,etc.).	5
Sample size	The study defines its sample (in terms of sample size).	5
Data collection method	The study specifies and thoroughly describes its data collection method (e.g. interviews, self-administrated questionnaires, etc.).	3
Questionnaire design	The study describes how the questionnaire was designed (e.g. the number of questions, type and wording of the questions, question sequence and grouping, translations, etc.).	4
Provisions for securing trustworthiness	The study describes the provisions made to secure trustworthiness (e.g. adoption of appropriate, well recognized research methods, examination of previous research findings, etc.)	3
Survey implementation		
Questionnaire evaluation	The study provides information on how the questionnaire was evaluated (e.g. through piloting, focus groups, in-depth interviews, statistical methods, etc.).	3
Questionnaire	The questionnaire of the study is available (e.g. attached to the report or included as an appendix, etc.).	3
Survey execution		
Media	The study describes how the survey was mediated (e.g. through mail, e-mail, telephone, web, etc.).	1
Execution time	The study specifies how long the survey was available to respondents.	1
Response burden	The study specifies how long the survey took to fill out by respondents.	1
Follow-up procedures	The study specifies the procedures taken in order to encourage response and prevent non-response.	2
Responses	The study provides information on the number of responses received.	3
Response rate	The study provides information on its response rate.	5
Survey analysis and packaging		
Assessment of trustworthiness	The study formally assesses its trustworthiness (e.g. through calculating measurement error, sample frame error, error of selection, non-response error, etc.).	5
Limitations	The study describes its limitations and threats to validity.	3

Chapter 10

Example Surveys

10.1 Survey on Communication in Software Engineering Projects

10.1.1 Research Objective

It aims to investigating the state of practice on Communications Characteristics of Agility (CCA) at Software Organizations from the point of view of software projects members. These Characteristics were extracted from a set of characteristics of agility identified through a large scale survey [9, 10] as relevant for introducing agility on software processes. They are: Being Collaborative, Feedback Incorporation, People Oriented, Being Cooperative, Reflection and Introspection and Self Organization. The definition of each characteristic can be found at [1]

10.1.2 Target Population

Professionals that recently worked in a software project at Swedish Software Organizations.

10.1.3 Sampling Design

It will be sent an invitation for each organization, asking for selecting recently concluded projects and recruiting their software practitioners on answering the questionnaire

Internal Questions

- How the Communication Characteristics of agility (CCA) have been applied in software projects?

- Is there any influence from the project attributes on the communication effectiveness?
- What is the relation between the project' effectiveness perceived by the team member and his/her opinion regarding the communication effectiveness?
- What are the gaps and challenges on improving the communication on software engineering projects?

Questionnaire- Attributes

- Company Name
- How many time ago your last concluded project was finished?
- Was this project performed in you current company?
- Team size?
- Your time allocation (part-time, full time)
- Are the team physically distributed? How? (2 or more rooms in the same building, 2 or more buildings, 2 or more countries)
- What was the project official language? (Swedish, English, Other)
- What was the followed software development process? (RUP, RUP Small, Scrum, XP, LSD, Kanbam, Crystal, Other)
- What was your main activity in the project? (programming, requirements engineering, testing, management, inspection, documentation, design, other)

Questionnaire- Opinion

Questions were derived from the concepts of each CCA described by Abrantes and Travassos Abrantes13 and adapted from the survey on Five Agile Factors presented at Stettina11 Each question was designed to be answer considering the following Likert Scale: totally disagree, partially disagree, partially agree, totally agree.

1. Project results
    - The project had attended to its original scope.
    - The project had attended to the customer needs.
    - The time of the project was well applied.

- The resources available was well applied.

2. Self Organization
   - Everyone was involved in the decision-making process.
   - Team members made important decisions without consulting other team member.
   - The team vision was well defined and presented.
   - The team was designed and redesigned according to its purpose.
3. People Orientation
   - The team took into account alternative suggestions in team discussion.
   - The team valued alternative suggestions.
   - Team members related to the tasks of individuals.
4. Being Collaborative
   - The role of everyone in the project was clear for me.
   - The roles needed on each activity that I participated have been applied in harmony, avoiding communication gaps.
   - The overall goal of the project was clear for me.
   - I didn't have difficulties on keeping in touch with my team partners.
5. Feedback Incorporation
   - I regularly comment on a co-worker‘s work.
   - The team gave feedback on all aspects of each others work.
6. Reflection and Introspection
   - Team keeps what works well in the development process.
   - The team improved the developments method when software development problems was identified.
   - Team met after each sub-project.
   - Team met after each iteration.
   - The project current status was easily accessible for all the project members.
7. Being Cooperative
   - It was easy to keep in touch with costumers' representatives.

- Customers' representatives actively worked in the projects.
- Customers' representatives regularly gave feedback to the project team.
- The project team was accessible to the costumer.

Bibliography

[1] J. F. Abrantes and G. H. Travassos, "Towards pertinent characteristics of agility and agile practices for software processes," *CLEI Eletronic Journal*, vol. 16, no. 1, p. Paper 5, 2013.

[2] S. Bennett, T. Woods, W. M. Liyanage, and D. L. Smith, "A simplified general method for cluster-sample surveys of health in developing countries," *World Health Stat Q*, vol. 44, no. 3, pp. 98–106, 1991.

[3] N. Bettenburg, S. Just, A. Schroter, C. Weiss, R. Premraj, and T. Zimmermann, "What makes a good bug report?" in *Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering*. ACM, 2008, pp. 308–318.

[4] H. N. Boone and D. A. Boone, "Analyzing likert data," *Journal of Extension*, vol. 50, no. 2, pp. 1–5, 2012.

[5] J. Boustedt, "A methodology for exploring students' experiences and interaction with large-scale software through role-play and phenomenography," in *Proceedings of the ACM Workshop on International Computing Education Research*, 2008, pp. 27–38, cited By (since 1996)0.

[6] M. Ciolkowski, O. Laitenberger, S. Vegas, and S. Biffl, *Practical experiences in the design and conduct of surveys in empirical software engineering*. Springer, 2003.

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

[8] R. M. de Mello, P. C. Silva, and G. H. Travassos, "Agilidade em processos de software: Evidências sobre características de agilidade e práticas ágeis," in *Brazilian Symposium on Software Quality*. Softex, 2014, pp. 308–318.

[9] R. M. de Mello and G. H. Travassos, "Would sociable software engineers obsverve better," in *Proceedings of 7th International Conference on*

Empirical Software Engineering and Measurement (ESEM). IEEE, 2013, pp. 279 – 282.

[10] R. M. de Mello, P. C. da Silva, P. Runeson, and G. H. Travassos, "Investigating probabilistic sampling approaches for large-scale surveys in software engineering," in *Proceedings of 11th Workshop on Experimental Software Engineering (ESELAW 2014)*, 2014.

[11] T. Dyba, "An empirical investigation of the key factors for success in software process improvement," *Software Engineering, IEEE Transactions on*, vol. 31, no. 5, pp. 410–424, 2005.

[12] S. Elo and H. Kyngas, "The qualitative content analysis process," *Journal of Advanced Nursing*, vol. 62, no. 1, pp. 107–115, 2008.

[13] W. Fan and Z. Yan, "Factors affecting response rates of the web survey: A systematic review," *Computers in Human Behavior*, vol. 26, no. 2, pp. 132 – 139, 2010.

[14] Y. Ghanam, F. Maurer, and P. Abrahamsson, "Making the leap to a software platform strategy: Issues and challenges," *Information and Software Technology*, vol. 54, no. 9, pp. 968–984, 2012, cited By (since 1996)4.

[15] P. I. Good, *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005, vol. 3.

[16] S. GROUP *et al.*, "The chaos chronicles," *The Standish Group International*, 2003.

[17] E. G. Guba, "Criteria for assessing the trustworthiness of naturalistic inquiries," *ECTJ*, vol. 29, no. 2, pp. 75–91, 1981.

[18] H. Haeger, A. D. Lambert, J. Kinzie, and J. Gieser, "Using cognitive interviews to improve survey instruments," in *The annual forum of the Association for Institutional Research New Orleans, Louisiana*, 2012, presented at the annual forum of the Association for Institutional Research New Orleans, Louisiana.

[19] T. A. Heberlein and R. Baumgartner, "Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature," *American Sociological Review*, vol. 43, no. 4, pp. pp. 447–462, 1978.

[20] H.-F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," *Qualitative Health Research*, vol. 15, no. 9, pp. 1277–1288, 2005.

[21] R. Johns, "Likert items and scales. Accessed on 13th March 2015 at: www.surveynet.ac.uk/sqb/datacollection/likertfactsheet.pdf."

[22] J. Joseph F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. Prentice Hall, 2009.

[23] S. K. Kachigan, *Statistical analysis: An interdisciplinary introduction to univariate & multivariate methods.* Radius Press New York, 1986.

[24] ——, *Multivariate statistical analysis: A conceptual introduction.* Radius Press, 1991.

[25] M. Kasunic, "Designing an effective survey," DTIC Document, Tech. Rep., 2005.

[26] L. Kish, *Survey Sampling.* Willey, 1995.

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

[28] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. E. Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, 2002.

[29] K. Krippendorff, *Content Analysis: An Introduction to Its Methodologyy.* SAGE Publications, 2012.

[30] M. S. Litwin, *How to Measure Survey Reliability and Validity.* SAGE Publications, Thousand Oaks, CA, 1995.

[31] M. K. Malhotra and V. Grover, "An assessment of survey research in pom: from constructs to theory," *Journal of operations management*, vol. 16, no. 4, pp. 407–425, 1998.

[32] B. F. Manly, *Multivariate statistical methods: a primer.* CRC Press, 2004.

[33] K. Moløkken-Østvold, M. Jørgensen, S. S. Tanilkan, H. Gallis, A. C. Lien, and S. Hove, "A survey on software estimation in the norwegian industry," in *Software Metrics, 2004. Proceedings. 10th International Symposium on.* IEEE, 2004, pp. 208–219.

[34] J. Moore, J. Pascale, P. Doyle, A. Chan, and J. K. Griffiths, *Using Field Experiments to Improve Instrument Design: The SIPP Methods Panel Project.* John Wiley & Sons, Inc., 2004, pp. 189–207.

[35] A. Pinsonneault and K. L. Kraemer, "Survey research methodology in management information systems: an assessment," *Journal of management information systems*, pp. 75–105, 1993.

[36] S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, *Methods for Testing and Evaluating Survey Questionnaires*. Wiley, 2004.

[37] L. M. Rea and R. A. Parker, *Designing and conducting survey research: a comprehensive guide*, 3rd ed. San Francisco: Jossey-Bass Publishers, 2005.

[38] L. Roberts, R. Lafta, R. Garfield, J. Khudhairi, and G. Burnham, "Mortality before and after the 2003 invasion of iraq: cluster sample survey," *The Lancet*, vol. 364, no. 9448, pp. 1857–1864, 2004.

[39] C. Robson, *Real World Research - A Resource for Social Scientists and Practitioner-Researchers*, 2nd ed. Malden: Blackwell Publishing, 2002.

[40] P. Rodríguez, J. Markkula, M. Oivo, and K. Turula, "Survey on agile and lean usage in finnish software industry," in *Proceedings of 6th International Symposium on Empirical Software Engineering and Measurement (ESEM)*. ACM, 2012, pp. 139–148.

[41] C. E. Sarndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, 1st ed. Springer, 1992.

[42] H. Sharp, M. Woodman, and F. Hovenden, "Tensions around the adoption and evolution of software quality management systems: A discourse analytic approach," *International Journal of Human Computer Studies*, vol. 61, no. 2, pp. 219–236, 2004.

[43] ——, "Using metaphor to analyse qualitative data: Vulcans and humans in software development," *Empirical Software Engineering*, vol. 10, no. 3, pp. 343–365, 2005.

[44] A. K. Shenton, "Strategies for ensuring trustworthiness in qualitative research projects," *Education for information*, vol. 22, no. 2, pp. 63–75, 2004.

[45] S. Siegel and N. J. Castellan, *Nonparametric statistics for the behavioural sciences*. McGraw-Hill, 1988.

[46] S. Stavru, "A critical examination of recent industrial surveys on agile method usage," *Journal of Systems and Software*, 2014.

[47] S. M. Sulaman, K. Wnuk, and M. Höst, "Perspective based risk analysis - a controlled experiment," in *Proceedings of the 18th International*

Conference on Evaluation and Assessment in Software Engineering, ser. EASE '14. ACM, 2014, pp. 47:1–47:10.

[48] S. K. Thompson, *Sampling*, 3rd ed. Wiley, 2012.

[49] U. van Heesch and P. Avgeriou, "Mature architecting-a survey about the reasoning process of professional architects," in *Software Architecture (WICSA), 2011 9th Working IEEE/IFIP Conference on*. IEEE, 2011, pp. 260–269.

[50] R. Van Solingen and E. Berghout, *The Goal/Question/Metric Method: a practical guide for quality improvement of software development*. McGraw-Hill London, 1999, vol. 40.

[51] M. D. White and E. E. Marsh, "Content analysis: A flexible methodology," *Library Trends*, vol. 55, no. 1, 2006.

[52] G. B. Willis, "Cognitive interviewing: A "how to" guid," Research Triangle Institute, Tech. Rep., 1999, reducing Survey Error through Research on the Cognitive and Decision Processes in Surveys: A short course presented at the 1999 Meeting of the American Statistical Association.

[53] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer, 2012.

[54] F. J. Yammarino, S. J. Skinner, and T. L. Childers, "Understanding mail survey response behavior a meta-analysis," *Public Opinion Quarterly*, vol. 55, no. 4, pp. 613–639, 1991.