CONCEPTUAL FRAMEWORK FOR SUPPORTING THE IDENTIFICATION OF REPRESENTATIVE SAMPLES FOR SURVEYS IN SOFTWARE ENGINEERING

Rafael Maiani de Mello

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Guilherme Horta Travassos

Rio de Janeiro
Março de 2016

CONCEPTUAL FRAMEWORK FOR SUPPORTING THE IDENTIFICATION OF REPRESENTATIVE SAMPLES FOR SURVEYS IN SOFTWARE ENGINEERING

Rafael Maiani de Mello

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Guilherme Horta Travassos, D.Sc.

_____

Prof[a]. Ana Regina Cavalcanti da Rocha, D.Sc.

_____

Prof. Toacy Cavalcante de Oliveira, D.Sc.

_____

Prof. Alessandro Fabricio Garcia, D.Sc.

_____

Prof. Rafael Prikladnicki, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2016

Mello, Rafael Maiani de

Conceptual Framework for Supporting the Identification of Representative Samples for Surveys in Software Engineering / Rafael Maiani de Mello – Rio de Janeiro: UFRJ/COPPE, 2016.

XIV, 138 p.: il.; 29,7 cm.

Orientador: Guilherme Horta Travassos.

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 115-124.

1. Survey research. 2. Conceptual framework. 3. Sampling. 4. Recruitment. 5. Representativeness. 6. Experimental Software Engineering. I. Travassos, Guilherme Horta II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título

To my beloved father Wilson *(in memorian)*

To Raquel and Miguel

Acknowledgments

My wife Fernanda is the great responsible by the success of this journey in the backstage. Due to her unconditional love and understanding, it has been pleasant and fruitful conciliating family, research and work. I thank God for her and my children. Fernanda, you always will be the great partner of my life!

I want to express my deepest sense of gratitude to my advisor Prof. Guilherme Horta Travassos. His daily examples of professionalism and detachment are inspiring. It is impossible to measure how much I learned with him during all these years. Indeed, more than an advisor, Guilherme became a best friend to me. Thank you, GHT!

I am also thankful to Prof. Per Runeson for opening the doors of SERG (Software Engineering Research Group, Lund University) to me. His advices were very important to reflect about the directions of my research, also opening my mind to new perspectives. At SERG, I could feel at home in a productive and exciting research environment, resulting in important collaborations even after the internship program.

I acknowledge my gratitude to Prof. Rafael Prikladnicki, Prof. Alessandro Garcia, Prof. Ana Regina Rocha and Prof. Toacy Oliveira to the participation in the Thesis' committee.

The support and advices of many other researchers were important not only to evolve my doctoral research but also to establish new research partnerships. In special, I am thankful to prof. Kathryn Stolee (North Carolina State University), prof. Martin Höst (Lund University) and prof. Claudia Werner.

It was an honor being a member of the ESE Group during the last five years. At ESE Group, I had found great research colleagues and had made many friends. I could work more closely with some of them (Jobson Massollar, Pedro Correa, Paulo Sergio Medeiros, Talita Ribeiro, Breno França, Rebeca Motta, Victor Vidigal, Verônica Vaz), but I want to thank all of them!

I am thankful to all professors, colleagues and staff from PESC that are continuously working to provide a high-level research program. In special, I would like to thank to Prof. Jano Moreira de Souza, Renata Mesquita, Eldanae Teixeira, Ana Prata, Solange Santos and Gutierrez da Costa. I also thank to the reception and support of all members from the SERG group and its technical staff.

Hundreds of Software Engineering researchers and practitioners spread in the world had devoted their time to participate in different investigations addressed to my Thesis. Probably most of them will never read this, but I acknowledge my gratitude to each one.

I would like thank also to the love and support of my dear mother Teresa Cristina, my grandmother Aurea and my sister Érika. I am also thankful to all my colleagues and students to the valuable opportunity of exchanging knowledge. I am also grateful to all my friends, in special Cônego Geraldo, Dr. Wilson, Renata, Luíza, João Baptista, Rogério, Reinaldo Marcos Ribeiro, Sérgio, Bruno, Denise, Lars Nilsson and Elaine.

Finally, I acknowledge my gratitude to CNPq and COPPETEC foundation by the financial support and to my colleagues at Banco do Brasil S. A. by the contribution to make possible my participation in the sandwich doctoral program.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

ARCABOUÇO CONCEITUAL PARA APOIAR A IDENTIFICAÇÃO DE AMOSTRAS REPRESENTATIVAS PARA SURVEYS NA ENGENHARIA DE SOFTWARE

Rafael Maiani de Mello

Março/2016

Orientador: Guilherme Horta Travassos

Programa: Engenharia de Sistemas e Computação

A representatividade das amostras utilizadas em estudos primários em Engenharia de Software é ainda um grande desafio, especialmente na condução de pesquisas de opinião (*surveys*). Este desafio inclui, entre outros, a identificação de fontes disponíveis para o estabelecimento de quadros de amostragem adequados, a caracterização de necessariamente indivíduos *(subjects*) e a necessidade de estimular a participação destes indivíduos. Apesar da importância das pesquisas baseadas em *survey* para a área, os poucos *guidelines* disponíveis para condução de *surveys* na Engenharia de Software raramente tratam destas questões. A pesquisa apresentada nesta Tese introduz um arcabouço conceitual para apoiar pesquisadores no estabelecimento sistemáticos de amostras representativas para *surveys* na Engenharia de Software. Atualmente em sua terceira versão, este arcabouço conceitual é composto por um conjunto de atividades para planejamento de *surveys* e tarefas projetadas para apoiar a instanciação de seus conceitos. Ele também oferece 40 recomendações para conduzir estas tarefas, derivadas de lições aprendidas na aplicação das versões anteriores da tecnologia e da literatura técnica. De acordo com os resultados dos estudos realizados, a versão apresentada nesta Tese possui maturidade suficiente para ser utilizado por pesquisadores interessados em planejar *surveys*.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

CONCEPTUAL FRAMEWORK FOR SUPPORTING THE IDENTIFICATION OF REPRESENTATIVE SAMPLES FOR SURVEYS IN SOFTWARE ENGINEERING

Rafael Maiani de Mello

March/2016

Advisor: Guilherme Horta Travassos

Department: Systems Engineering and Computing

Samples representativeness in Software Engineering primary studies is still a great challenge, especially when conducting opinion surveys. Such challenge includes among other issues the identification of sources available for establishing adequate sampling frames, the characterization of survey subjects (necessarily individuals), and the need for stimulating subjects' participation. Despite the importance of survey research to the field, the few guidelines for conducting surveys in Software Engineering available in the technical literature barely address these issues. The research presented in this Thesis introduces a conceptual framework structured for supporting researchers on systematically establishing representative samples for surveys in Software Engineering. Currently in its third version, such framework has evolved through experimentation since its first one. The conceptual framework is composed by a set of survey planning activities and tasks designed for supporting the instantiation of its concepts. It also provides 40 recommendations for conducting its tasks, derived from lessons learned on applying the previous versions of the technology and from the technical literature. According to the results of the performed studies the current version of the conceptual framework has enough maturity to be used by researchers interested on planning surveys.

INDEX

INDEX OF FIGURES

INDEX OF TABLES

1 Introduction

In this chapter we introduce the problem and the context involved in this Thesis, as well as the research questions supporting the investigation. Furthermore, we establish the objectives to be accomplished in order to answer the research questions and how they will be performed through an evidence-based methodology.

1.1. Context

Empirical Software Engineering practices have contributed in the last decades to reduce the empiricism in which Software Engineering (SE) technologies have been introduced in the field, leading the voice of evidence to the state of practice. Since the publication of the seminal paper *"The experimental paradigm in Software Engineering"* (BASILI, 1993), several investigations have been conducted in many SE disciplines such as *requirements engineering* (KARLSSON, WOHLIN and REGNELL, 1998; RICCA *et al.*, 2009; WNUK, GORSCHEK and ZAHDA, 2013), *software inspection* (TRAVASSOS *et al.*, 1999; AURUM, PETERSSON, and WOHLIN, 2002; BIFFL, GRÜNBACHER, and HALLING, 2006), *software testing* (KUHN, WALLACE and GALLO, 2004; JURISTO, MORENO and VEGAS, 2004; RAFI *et al.*, 2012) and *software maintainability* (CHEN and HUANG, 2009; FERNÁNDEZ-SÁEZ *et al.*, 2014; HANENBERG *et al.*, 2014), among others.

If in the early years of empirical SE the research efforts were concentrated on conducting controlled experiments (SJØBERG *et al.*, 2005), currently many qualitative research methods have been applied to support in depth investigations, including *case study* (RUNESON AND HÖST, 2008), *action research* (FÆGRI, DYBÅ and DINGSØYR, 2010) and *focus group* (RODRÍGUEZ *et al.*, 2013). In this sense, DYBÅ, SJØBERG and CRUZES (2012) argue that SE researchers should immerse in the context by increasingly conducting high-quality qualitative studies focused in relevant phenomena from the practice. However, BEN SHNEIDERMAN (2013) argued in a keynote speech that doing *High Impact Research* requires researchers to avoid prioritizing both *basic research* - typically supported by controlled experiments - and *applied research*- typically supported by qualitative studies. Instead, we should recognize benefits from both

activities, harmonically blending them for delivering relevant contributions to both academy and industry.

In this context, the versatility of the *survey method* on supporting basic and applied researches can be observed. The survey method is an observation strategy useful to collect information regarding events and/or phenomena, identifying *trends* and *consensus* in a specific research context (LEVIN, FOX and FORDE, 2012). PINSONNEAULT and KRAEMER (1993) classify as *survey research* those conducted to advance scientific knowledge, aiming at producing quantitative descriptions of some aspects of the studied population by asking people structured and predefined questions.

When properly conducted, *opinion surveys*[1] (hereinafter simply called as *surveys*) allow researchers to perform descriptive and large scale investigations without the rigorous level of control required by controlled experiments, supporting the characterization of *knowledge*, *attitudes* and/or *behaviors* from different groups of individuals (KASUNIC, 2005) through the generalization of findings from a fraction of the population to the whole population. In SE research, surveys have been commonly adopted for different research goals, such as *mapping the state of practice* (CIOLKOWSKI *et al.*, 2003; STAVRU, 2014), *establishing baselines for investigating new fields* (STETTINA and HEIJSTEK, 2011), *gathering of opinion regarding SE technologies and practices* (FERNÁNDEZ-SÁEZ *et al.*, 2013), among others.

1.2. Problem Definition

BACKSTROM and HURSH-CÉSAR (1981) presents the following important characteristics of a survey: it should be *systematic*, *impartial*, *representative, theory-based*, *quantitative* and *replicable*[2]*.* It means that researchers should select elements of the population that together are *representative* of the problem under study without prejudice or preference (impartial), being guided by relevant principles of human behavior and mathematical laws (theory-based), assigning numerical characteristics of human behavior in ways that allows uniform interpretation of these characteristics (quantitative). Finally, a survey should follow a specific and formal set of steps (systematic) so that different researchers following the same methods in the same

[1]i.e., surveys in which necessarily the study subject is the individual.

[2]In SE literature, “replication” is a term commonly used to characterize experimental studies, but it is not a consensus to other study types. Thus, we use in this document the term “re-execution” when referring to surveys replication.

conditions can get essentially the same results (replicable). Although the survey method is one of the most frequent research methods applied to Empirical Software Engineering (ESE), the external validity of SE surveys are often impacted by the use of convenience for sampling (DE MELLO *et al*., 2015). The previous experience of the Experimental Software Engineering Group on investigating and conducting surveys indicated that even after exhaustive effort on sampling and recruitment activities, SE survey executions frequently fail to be *impartial* and *representative*. Moreover, since such effort is typically grounded in *ad-hoc* activities, SE surveys also fail to be completely *systematic* and, consequently, *replicable*. For instance, CONRADI *et al.* (2005) evaluated a set of surveys on component based SE and concluded that most of them do not let clear how their samples were established. A similar conclusion was obtained by STAVRU (2014) when evaluating surveys conducted in academy and industry to investigate the using of agile methods in software organizations. The author states that it is not possible to assure to which extent the results obtained by most of the analyzed surveys could be considered valid.

More recently, we identified through a structured review[3] (DE MELLO and TRAVASSOS, 2015) that most of the surveys published in the International Conference on Evaluation and Assessment in Software Engineering (EASE) and the International Symposium on Empirical Software Engineering and Measurement (ESEM) conferences since 2005 were supported by non-representative samples established by convenience, including *business partners*, *local students* and *researchers' personal networks*.

One can see that not only challenges on characterizing the diverse SE research contexts contribute to the observed scenario (DYBÅ, SJØBERG and CRUZES, 2012), the business nature of SE also do, typically restricting the access to large data sets in the field, such as information about organizations' professionals and projects. As a comparison, surveys from other fields such as *social sciences*, *education* and *nursing* are commonly supported by country-wide sampling frames composed of large sets of *citizens/households*, *students/classes/schools* and *patients/hospitals*, respectively. Thus, one key issue on establishing representative samples to support SE surveys relies

[3]As defined in (eSEE 2015):*"…characterize studies in which researchers do not intend to exhaustively characterize a field of interest neither to gather all published evidence available to understand a phenomenon. It allows to systematically map publications from a research field in a reduced but recognized relevant subset of sources (eventually from a specific period), such as relevant journals and conferences regarding a specific research field."*

on identifying relevant and accessible sources from which adequate sampling frames can be organized. In this sense, technical literature reveals a few examples of the use of alternative sources available in the Web, such as professional social networks (KANIJ, MERKEL and GRUNDY, 2011; JOORABCHI, MESBA and KRUCHTEN, 2013) and discussion groups (NUGROHO and CHAUDRON, 2008) but such use is typically addressed to enlarge the number of respondents rather than to provide samples' representativeness.

In the context of our research, a *representative sample* consists of a subset of units, randomly retrieved from a *heterogeneous* population from the point of view of the survey target audience attributes (DE MELLO *et al.*, 2015). Such definition addresses three representative samples' quality properties out of the nine described by KRUSKAL and MOSTEELER (1979): (1) *specific sampling method* (probabilistic sampling), (2) *populations' heterogeneity coverage* and (3) *representative as typical*, with respect to certain known population attributes, such as *gender*, *age* and *income*. Thus, the survey population heterogeneity and the sampling method adopted should determinate the survey sample size, not vice versa.

In this sense, it is important to point out different target audiences and research objectives will demand more/less effort on sampling, which is not necessarily related with the population size. For instance, a survey specifically designed to a local organization may need stratifying its population by different departments and roles; while an international large-scale survey with SE professionals may not demand stratification efforts.

In addition, establishing representative samples may not be sufficient to assure results representativeness since the participation in surveys is commonly voluntary. Hence, a survey plan should also establish how to systematically encourage responses and prevent non-responses (STAVRU, 2014). In this context, the investigation of 52 surveys published from 2005 to 2014 at ESEM and EASE proceedings allowed us to observe that surveys' subjects are often invited using different methods and instruments to the same study, whereas *persuasive factors* (SMITH *et al.*, 2013) are eventually applied to encourage their participation.

1.3. Research Questions

Based on the presented issues regarding survey research in SE, the following research questions emerged:

RQ1. How to identify and assess potentially relevant sources of populations available for conducting surveys in SE?

RQ2. How to deal with the limitations on retrieving relevant information from these sources?

RQ3. How to characterize samples for surveys in SE?

RQ4. How to stimulate participation for surveys in SE?

RQ5. How to systematize all the sampling and recruitment activities in order to make them repeatable?

Adapting a research method to a new field is not simple. Although the largely use of survey research in the field, we claim that such adaptation should be better supported. In this context, an investigation over known guidelines for conducting surveys in SE (LINÅKER *et al.*, 2015) identified that they provide insufficient guidance to answer such questions. They commonly reproduce concepts and practices from the survey research principles in general, even reporting challenges on the survey planning in SE without providing adequate orientation on how to overcome them.

1.4. Research Objective

The research objective of this thesis is *to establish a conceptual framework to support researchers to conduct their survey planning activities by guiding the systematic identification of representative samples for surveys in SE.*

In this sense, it is important to emphasize that the scope of the technology proposed in this thesis does not include the whole survey process and does not even include all survey planning activities, focusing on providing guidance to mitigate the external threats to validity often observed in SE surveys regarding samples' representativeness.

1.4.1. Expected Contributions

The main expected contribution of the present research is *to improve the quality of SE survey plans in the context of sampling and recruitment activities through establishing a set of concepts, activities/ tasks and an initial set of useful recommendations grounded on specific SE research issues and on general survey practices*. Other contributions are also expected in the following research topics/ issues:

- *Improving the quality of large-scale experiments in SE,* due to the similarity of the issues regarding samples' representativeness observed in both research methods (DE MELLO *et al.*, 2015b);
- *Disseminating survey re-execution practice in the field*, through re-executing and analyzing SE surveys from diverse research topics;

– *Providing guidelines for characterizing the context of surveys in SE*.

Regarding the last item, JOHNS (1991) distinguishes between two types of research context: *substantive* and *methodological*. The methodological context refers to detailed information about the research study while the substantive context stands for the context individuals or groups face. The substantive context can be characterized through omnibus (broad) and discrete perspectives. GRIFFIN *et al*. (2007) observe that the omnibus perspective can be considered the lens from which the variables of the discrete perspective can be observed. Thus, taking into account the large variability of context in SE research, DYBÅ, SJØBERG and CRUZES (2012) encourage SE researchers to take an omnibus perspective to characterize the substantive context of their studies through answering the following five questions: *Who?, What?, Why?, When?* and *Where?*

Figure 1-1 highlights the potential contributions of the presented research to the context characterization of surveys in SE. Besides the expected contribution of our research to survey planning (methodological context), we also expect contributions to the characterization of surveys' substantive context, especially from an omnibus perspective. Based on a given phenomenon to be investigated (What?), the conceptual framework guides the identification of a representative sample (Who?) and how to recruit such sample by stimulating subjects participation (Why?). Moreover, the systematization provided by the conceptual framework may also contribute to reflecting on *when* and *where* the survey should be executed. Finally, since the conceptual framework provides some guidance to characterize subjects, we also expect that it can be useful for characterizing the *social dimension* of SE surveys.

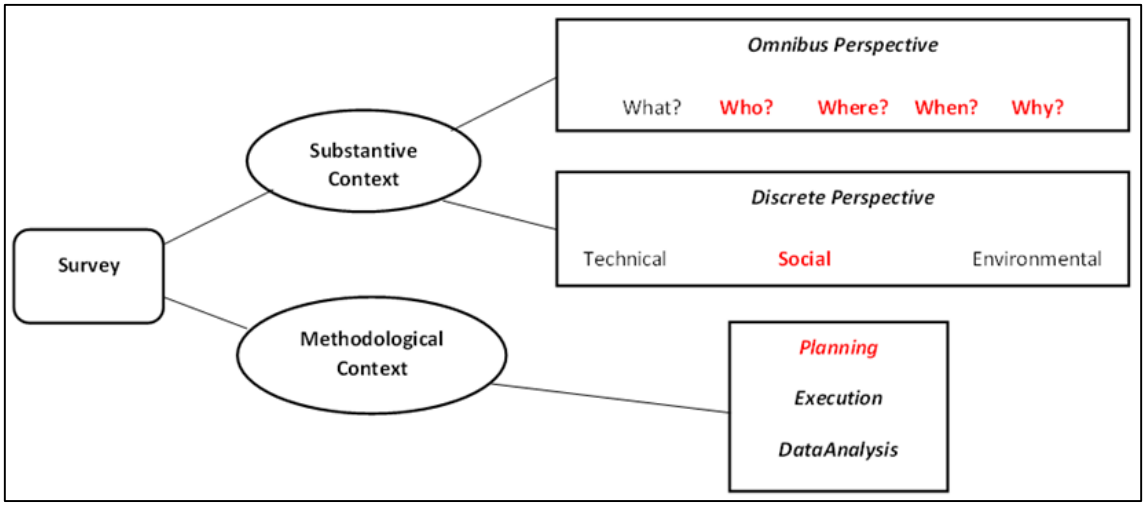


Figure 1-1. Expected conceptual framework contributions to the characterization of SE research contexts (in red).

1.5. Related Research and Academic Context

The presented (meta) research is inserted in the context of the ESE group supervised by Professor Guilherme Travassos since 2001, devoted to promote empirical research in the field by *"studying and researching new models to support the planning, execution and packaging of SE empirical studies"* (ESE, 2015). The presented research is also inserted in the context of a broader investigation conducted by the ESE group in partnership with the Software Engineering Research Group (SERG) at Lund University (Sweden) to study the guidelines to conduct surveys in SE. Co-supervised by Professor Per Runeson, the author of the presented thesis had participated in an internship program, partially conducting his doctoral activities located at SERG. Among the research activities performed at SERG, the author participated in a survey course where guidelines to conduct surveys in SE were investigated (LINÅKER *et al.*, 2015). Also, we designed and started a Systematic Literature Review (SLR) to investigate additional guidelines with the SERG.

The opportunity of re-executing surveys from different research topics investigated by the ESE Groups allowed us to strengthen evidence regarding the findings obtained in the original executions. Such topics include *the introduction of agility in software processes* (ABRANTES and TRAVASSOS, 2013); *requirements effort influence factors* (VAZ, 2013) and *guidelines for conducting simulation based studies in SE* (FRANÇA and TRAVASSOS, 2015). We have also been collaborating with Professor Kathryn Stolee (North Carolina University) in the replication of online experiments (DE MELLO, STOLEE and TRAVASSOS, 2015) and providing the conceptual framework to be applied in the planning of new surveys conducted by other research groups. Such experiences allowed us to teach survey research classes from the Experimental Software Engineering course offered at COPPE/UFRJ (2014, 2015).

Finally, it is important to point out the receptivity of the Empirical Software Engineering community to our research ideas and concerns, which we could observe through participation in different venues in which research papers related with the presented thesis were presented (EASE 2013; ESEM 2013, 2014, 2015; ESELAW 2014, 2015; ELA-ES 2015).

1.6. Research Methodology

We originally have planned SE quantitative studies (experiments and quantitative surveys) as the scope of this thesis. Inspired by crowdsourcing classes at COPPE/UFRJ (2012, Professor Jano Moreira de Souza), we started investigating the potential

contributions of crowdsourcing technologies to support the enlargement of samples in SE quantitative studies (DE MELLO and TRAVASSOS, 2012). Then, we investigated the lack of external validity of such studies and the limitations to replicate them. Through the adaptation of concepts from *food chains* (Ecology) we structured the concept of *experimental chains* to characterize the current scenario of quantitative studies' replication in SE and to discuss alternatives to evolve such scenario (DE MELLO and TRAVASSOS, 2013). In this sense, we argued that most of the effort to conduct and replicate quantitative studies in SE is wasted due to the frequent use of convenience samples, restricting significantly the renewal of *energy* (evidence) in the field.

Then, taking into account the current and potential benefits of survey research to SE and the issues introduced in Section 1.2, we decided to change the scope of the research to surveys, following the research activities presented in Figure 1-2. The figure identifies the subset of research activities conducted until the qualification exam (presented in July, 2014) and the main research activities conducted during the internship at Lund University.

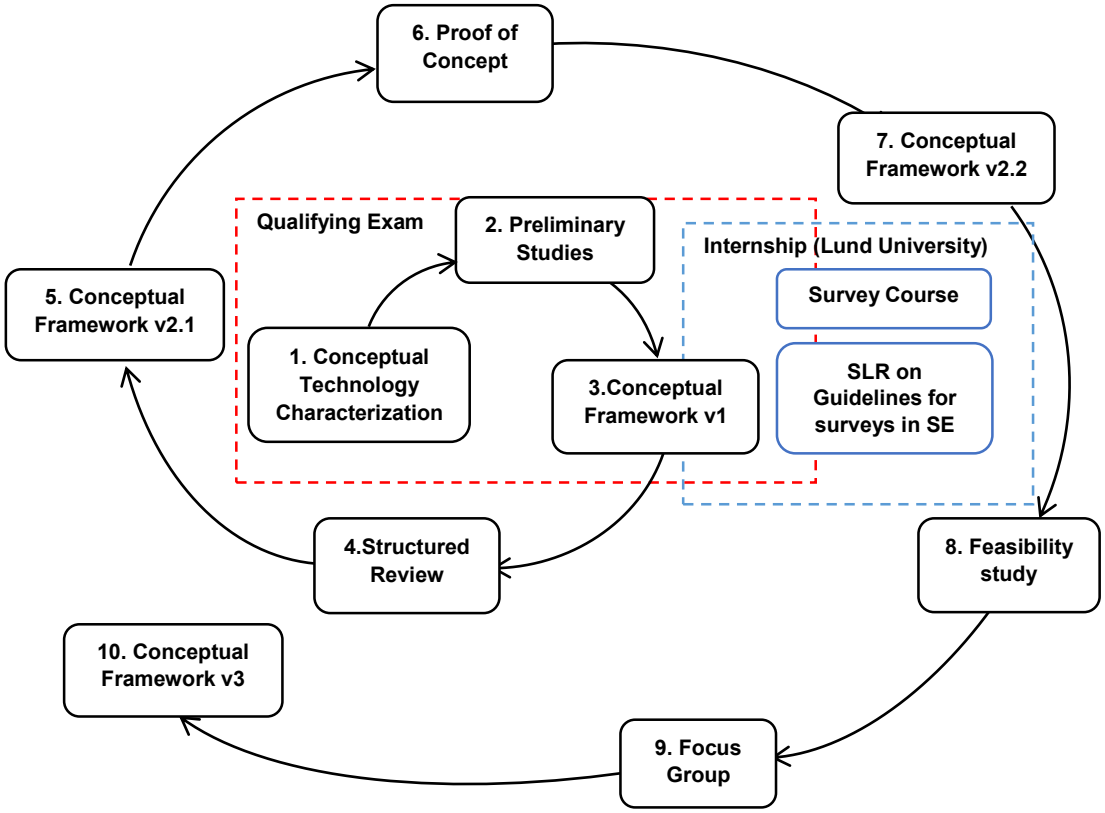


Figure 1-2.The research methodology.

Taking into account a scenario in which suitable sampling frames may be established through systematically searching for populations available in generic

sources, the Systematic Literature Review (SLR) concepts inspired us to perform the *conceptual technology characterization* (1). The concepts defined in the step 1 were then applied to support the conduction of *preliminary studies* (2) in which three distinct surveys planned by researchers from the Experimental Software Engineering (ESE) group at COPPE/UFRJ (VAZ 2013; ABRANTES and TRAVASSOS 2013; FRANÇA and TRAVASSOS, 2015) were re-executed. After applying different sampling designs over the sampling frames obtained from professional social networks, we evidenced that such re-executions allowed us to obtain more heterogeneous samples composed by more experienced subjects in the survey topic than the original executions. The description of such preliminary studies and discussions regarding their findings can be found in (DE MELLO and TRAVASSOS 2013b; DE MELLO, DA SILVA and TRAVASSOS, 2014a[4], 2014b, 2015).

Based on the lessons learned when conducting those studies, a *first version* of the conceptual framework was designed (3), available at (DE MELLO *et al.*, 2014c)[5]. As mentioned before, sampling issues involved in the presented investigation scope are not only observed in SE surveys but also in SE large-scale experiments. Consequently, the first version of the conceptual framework was used in the context of replicating an online experiment regarding Java code search (STOLEE, ELBAUM and DWYER, 2015). As a result, we observed that the effective sample obtained by instantiating the framework (composed by *LinkedIn* members from its most populous Java programmers group) was significantly more heterogeneous and more experienced in the research topic than the effective sample obtained in the original execution (Amazon's Mechanical Turk workers). The description of this study can be found in (DE MELLO, STOLEE and TRAVASSOS, 2015).

Aiming at providing guidance to other researchers when applying the conceptual framework, a *structured review* (4) was conducted over EASE and ESEM proceedings (electronically available at IEEE/ACM) on how sampling activities have been conducted in SE surveys (DE MELLO and TRAVASSOS, 2015). The same set of proceedings was then used to investigate how subjects have been recruited to SE surveys and the effect of using different sampling designs/ persuasive factors over subjects' participation (DE MELLO and TRAVASSOS, 2016). So, our experience with conducting the previous empirical studies led us to evolve the framework to a *second version* (5) (DE MELLO

[4] The work received the first best paper award at ESELAW 2014
[5] The work received the best short paper award at ESEM 2014.

and TRAVASSOS, 2015b) including activities, tasks and recommendations to guide the instantiation of the conceptual framework concepts.

A first release of this second version was submitted to a *proof of concept* (6) and a doctoral student at COPPE/UFRJ (external to our research group) was invited to use the conceptual framework to plan an actual survey. Based on its findings, a few improvements were made in the framework documentation and generated a *new release of the second version* (7). Such second release (v2.2) was then empirically evaluated through a *feasibility study* (8) in which the conceptual framework acceptance (usefulness, easiness of use, intention to use) and the effect of its use on the quality (thoroughness) of survey plans were investigated. Based on the findings from this study, a *focus group* session (9) was then conducted with the same subjects from the feasibility study to better understand whether the framework recommendations could contribute to plan a survey in SE (DE MELLO and TRAVASSOS, 2016). Finally, the findings from the previous empirical studies allowed us to evolve the conceptual framework to a *third version* (10).

1.7. Thesis Organization

This thesis is organized in six more chapters. Chapter 2 introduces the survey process and discusses related works. Chapter 3 summarizes the research steps conducted before the establishment of the second version of the conceptual framework. Section 4 presents the conceptual framework v2.2, empirically evaluated through the studies presented in Section 5. Section 6 introduces the third version of the conceptual framework. Conclusions and future work are presented in Section 7.

2 Related Work

In this chapter, we briefly introduce the typical survey process and the main statistical concepts related to it. Guidelines to conduct surveys found in the technical literature are also discussed.

2.1. Introduction

Samples' representativeness is a challenge not only in SE research. Throughout the last decades, fields with different maturity levels in survey research have been discussing how to improve survey samples' representativeness, such as *Marketing* (GOETZ, TYLER and COOK, 1984), *Health* (BRAITHWAITE *et al.*, 2002) and, closer to SE field, *Information Systems* (PINSONNEAULT and KRAMER, 1993). PINSONNEAULT and KRAMER (1993) analyzed 122 survey-based studies in Management Information Systems (MIS), reported in major MIS journals between 1980 and 1990, identifying than 70% of the studies used a convenience sample or did not report the sampling procedure. Also, the surveys which were analyzed frequently do not follow systematic procedures for sampling and their execution often results in poor participation rates. Therefore, the authors proposed some recommendations addressing the major problems observed, such as the following:

- improving the systematization of sampling procedures, investing efforts to adopt probabilistic sampling designs;
- keeping the survey questionnaire as short and simple as possible;
- taking into account the sample peculiarities when generalizing the survey findings, and;
- improving participation rates by getting endorsement from well-known professional associations and from top managers.

One can see a similarity among issues reported by Pinsonneault and Kramer 23 years ago in the context of Information Systems research and those exposed in the introduction of this thesis. Actually, we did not find a similar comprehensive investigation in the context of SE research.

During a survey course conducted in 2014 at SERG (Section 1.5), a set of SE papers were discussed and analyzed, hereinafter called *"survey guidelines"*. Such

guidelines include *methodological works*, i.e. publications explicitly devoted to present the survey method (PFLEEGER and KITCHENHAM, 2001; KASUNIC, 2005; KITCHENHAM and PFLEEGER, 2008) and *experience reports*, i.e. publications reporting surveys in SE but also discussing methodological issues and/or introducing guidelines (CIOLKOWSKI *et al.*, 2003; CONRADI *et al*; 2005; PUNTER *et al.*, 2005; STAVRU et al., 2014). Then, additional guidelines were also searched in the preliminary results from a comprehensive SLR conducted with the SERG. However, except by JI *et al.* (2008) and SMITH *et al.* (2013), no additional publications addressed to our research scope were found. Indeed, regardless of the frequent use of survey research in SE, we found few guidelines addressed to answer the thesis's research questions. Before discussing the content of such guidelines in Section 2.3, this chapter briefly introduces the survey process in Section 2.2, presenting the main statistical concepts related to it.

2.2. The Survey Process

Figure 2-1 shows the survey process followed in our investigation, presented by Kasunic's (2005) technical report, published by the Software Engineering Institute (SEI). The survey process described in Figure 2-1 can be summarized as follows: once the *research objective* is established, it should be characterized who is able to respond the survey (*characterize the target audience*) and then *design a sampling plan* to obtain sample from the survey population available. Next, researchers should *design and write the survey questionnaire*, testing it through a *piloting test questionnaire* with members from the target audience. Then, the survey should be executed (*distribute the questionnaire*). After the survey execution, its results should be finally *analyzed* and *reported*. One can see that Kasunic's four first steps are addressed to *survey planning* activities, while the second and third ones (*characterize the survey audience*/ *design the sampling plan*) are addressed to our research scope.

It is important to point out that we studied the survey process in the survey literature in general, identifying different arrangements of steps with different flows. For instance, the survey steps introduced by GROVES *et al.* (2009) support the *design of the survey questionnaire* concurrently with the *sampling design*. However, we argue that information regarding the survey population available could influence the composition of the survey questionnaire, especially taking into account the already mentioned lack of adequate sampling frames to support SE surveys. In addition, we observed that *planning the survey recruitment* is a key issue discussed in survey research (GROVES *et al.*, 2011; COHEN, MANION and MORRISON, 2013), although is not represented as a survey step by (KASUNIC, 2005). Such planning include, for instance, *establishing the*

recruitment message, *describing the recruitment procedures* and *characterizing eventual compensation that should be given to participants* (COHEN, MANION and MORRISON, 2013; SMITH *et al.*, 2013).

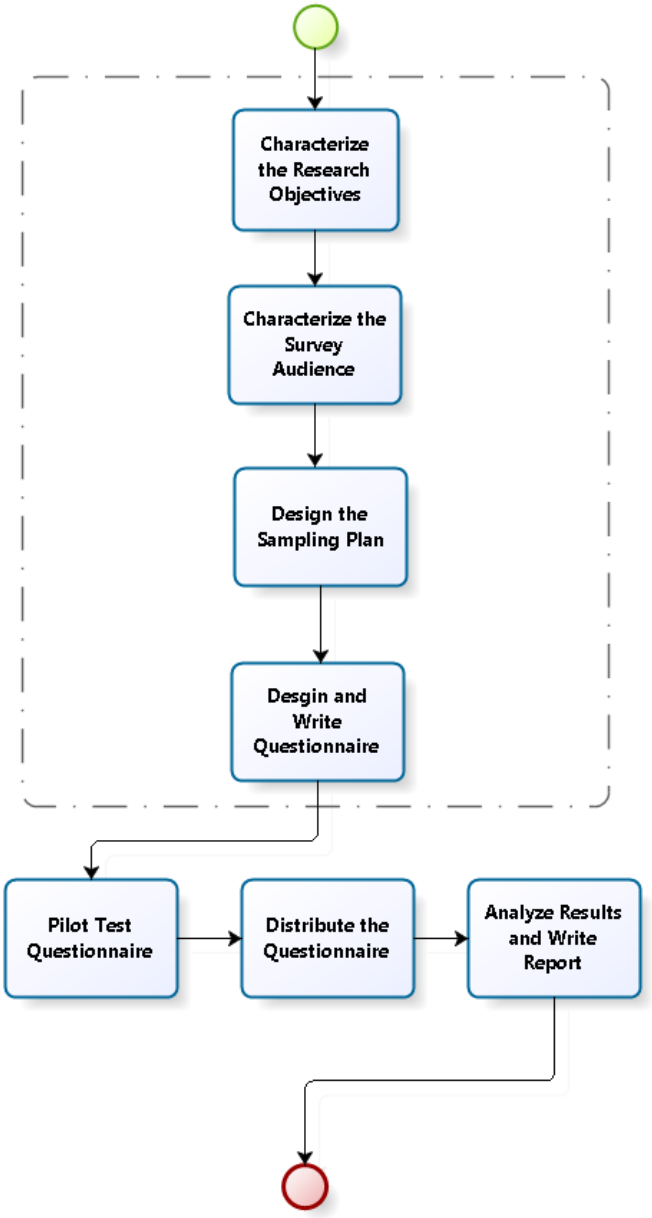


Figure 2-1. The survey process, based on (KASUNIC, 2005)

In order to understand the survey process it is important to first clarify the statistical concepts related to it. Table 2-1 presents the main statistical concepts addressed to sampling activities and adopted in our research with examples from the field. Apart from *unit of observation (subject)* and *unit of analysis,* all of these concepts were used since the beginning of our research.

Table 2-1. Statistical concepts used in the research, with examples.

Concept	Description	Examples in SE surveys
Target Audience[6]	Consists on a set of units that could be covered in a survey (GROVES et al., 2007), being such units composed by *individuals* or *groups of individuals*. The establishment of a survey target audience try to answer who can best provide the information needed in order to achieve the research objective	- Testing professionals - SE researchers - Java programmers
Unit of Observation (subject)	The minimum component from which data can be retrieved and analyzed from a study (HOPKINS, 1982).	The individual is always the unit of observation in opinion surveys.
Unit of Analysis	Consists on the major entity used for analyzing the study (Hopkins, 1982). In other words, unit of analysis is in the level at which the researchers *pitch the conclusions* while the unit of observation is in the level at which researchers *collect data*.	The individual is commonly also the survey unit of analysis, but groups of individuals such as organizations and project teams can be also used.
Population	Set of units from the target audience having a chance to be selected to compose the survey sample (GROVES *et al.,* 2009)	- Students from a Software Testing course - Audition from ICSE' 2015 - Java programmers working at Google
Sampling Frame	Listing of units from the target audience from which one or more samples can be retrieved (GROVES *et al.,* 2009). In the other words, it is the listing in which the survey population is available	- List of Students from a Software Testing course - List of ICSE' 2015 attendants - List of Java programmers employed at Google

[6] *Target audience* is also commonly known as *target population* in survey literature (GROVES et al., 2011; COHEN, MANION and MORRISON, 2013). However, we opted by using "*target audience"* in order to avoid ambiguous interpretation with the *frame population*, also known as only *population*.

Sampling	Process from which a *sample* is extracted from the sampling frame to support a survey execution (GROVES *et al.*, 2009). Sampling is typically needed when the effort involved on selecting all elements (*census*) composing a sampling frame is prohibitive, or even when the selection of all sampling frame could bring a side effect on the statistical power, introducing hypersensitivity to the sample (HAIR *et al*., 2009).	- All students from a Software Testing Course (census) - 383 ICSE 2015 attendants randomly selected (simple random sampling sampling) - A subset of experienced java programmers selected by the researchers based on their profiles (judgment sampling)
Sampling design	Describes how a specific sample size should be obtained from a sampling frame. In probabilistic sampling designs, all units from the study population have the same probability to be selected; while in non-probabilistic sampling designs don't (THOMPSON, 2012). *Accidental sampling, quota sampling, judgment sampling, snowball sampling* are examples of non-probabilistic sampling designs, while *simple random sampling, clustered sampling, stratified sampling* and *systematic sampling* are examples of probabilistic sampling designs.	

Regarding the characterization of the survey *target audience*, KASUNIC (2005) suggests considering questions such as the following:

- How many people are there in the population we are studying?
- What are their jobs and responsibilities?
- What is the most common education level?
- What relevant experience do they possess?
- What technical abilities do they possess?
- What is the age range of the respondent population?
- Do we anticipate that they would have difficulty in using a questionnaire that is: mailed to them, completed using a computer with Internet access or handed to them?
- What can we assume about their knowledge of the domain we are studying in the survey?
- How do they speak about the domain under study? Do they have a special language they use to talk about the problem space?
- Will they be motivated to complete the questionnaire? Or will they be resistant?
- How much of their time can we assume they will spend completing the questionnaire?

One can see that the consistency between concepts presented in Table 2-1 is essential for the success of the survey execution. For instance, if the survey target audience is not adequate to its research objective, all the research effort could be lost. In the same way, establishing a sample that is partially or totally out of the survey target audience could drive researchers to a misinterpretation of the survey results. Hence, sampling frames suitable to the survey target audience should be identified. From the point of view of its structure, an *ideal sampling frame* should present the following set of characteristics (SÄRNDAL, SWENSSON and WRETMAN, 1992):

1. All the elements have a unique logical/numerical identifier
2. All the elements can be retrieved and relevant information from them are available
3. The sampling frame is organized in a logical and systematic fashion
4. The sampling frame has additional information regarding its units
5. All the elements of the target audience are present in the sampling frame
6. All the elements of the target audience are present only once in the sampling frame
7. No element outside of the target audience is present in the frame

8. The data is up-to-date

The authors classify the first two characteristics as *essential* whereas the other characteristics are classified as *desirable*. The desirable characteristics have a potential contribution to providing sample representativeness and reduction of operational efforts to perform sampling activities. SÄRNDAL, SWENSSON and WRETMAN (1992) state that some appropriate investigations could not be carried out due to the lack of adequate sampling frames, while other investigation results remain inconclusive due to the convenience of the sampling frames used. However, it is common to observe the hazardous use of *convenience* for obtaining high participation in SE surveys, which includes sampling frames composed by *co-workers, colleagues*, *partners* and *clients*. In other cases, even a sampling frame is not established, commonly due to indirect recruitment. It happens, for instance, when subjects are indirectly recruited through generic messages published in open discussion groups or with crowdsourcing tools and, this way researchers are not able to control who are the individuals that could access them.

Regarding the sampling design, KASUNIC (2005) states that *"It is not acceptable to draw a non-probabilistic sample simply because you are unwilling to spend the effort, time, or money required to draw a true probability sample"*. Indeed, when generalizations from the sample are intended to be applied to the population, a probabilistic sampling design such as the commonly used *simple random sampling* must be used. In this sense, the s*ample size* of the survey should be calculated in function of the aimed *confidence interval* and *confidence level*. *Confidence level* is an index of how sure we can be that the survey responses will lie in a given variation range, i.e., a specific *confidence interval* (COHEN, MANION and MORRISON, 2013). It is commonly established a confidence level of 95% or 99% while the following formula can be used to calculate the sample size (KUPPER *et al.*, 1989):

$$SS = \frac{Z^2 \times p \times (1-p)}{c^2}$$ (1), where:

- *SS= sample size*
- *Z= Z-value, established through a specific table (Z=2.58 for 99% of confidence level, Z=1.96 for 95% of confidence level*
- *p= percentage selecting a choice, expressed as decimal (0.5 used as default for calculating sample size, since it represents the worst case).*
- *c= desired confidence Interval, expressed in decimal points (Ex.: 0.04).*

For the calculation of a sample size based on a *finite* population with a *pop* size, the following correction formula should be applied over *SS*:

$$SS_f = \frac{SS}{1+\frac{SS-1}{pop}} \quad (2)$$

For instance, considering a population composed of 10.000 individuals in which is acceptable that the observed results could vary ±10 points (confidence interval) for a confidence level of 95%, a sample composed of at least by 95 individuals is needed. However, in this example, sampling 95 individuals is only recommended if you can be sure that all subjects will effectively participate in the study. Since participation in surveys is often voluntary, higher sample sizes should be established to mitigate the impact of the *participation rate*. For instance, if previous experience indicates that only 20% of the subjects tend to participate effectively in the exemplified study, it can be considered a good practice to recruit a sample size five times higher than the calculated sample size (475).

2.3. Guidelines for Conducting Surveys in Software Engineering

In a series of five short papers presenting the principles of survey research, PFLEEGER and KITCHENHAM (2001) devoted one of them to discuss the issues regarding the survey design, emphasizing that researchers must keep in mind the following aspects when sampling from a population: *avoidance of bias, appropriateness* and *cost-effectiveness*. In this sense, the researchers recommended the use of Web technologies to reduce recruitment costs and presented some challenges to improve the response rates, pointing out that respondents should be *able to answer the survey questions* as well as be *willing* and *motivated* to answer them.

PFLEEGER and KITCHENHAM devoted, then, a second paper (2002) to introduce the principles of *population* and *sampling*, emphasizing that is not possible to sample from a population if such population is unknown. First, the survey *target audience* should be derived from the *research objective*. Then, a list composed of a subset of elements from the audience should be established, comprising the survey *sampling frame*. Finally, a representative *sample* should be extracted from this sampling frame. Afterwards, the researchers present the most common designs to perform sampling (probabilistic/ non-probabilistic), also introducing a statistical formula to calculate the survey sample size.

The recommendations made in the survey research papers were, then, compiled and extended in a book chapter (KITCHENHAM and PFLEEGER 2008), in which the researchers introduced the following set of strategies to improve the subjects' participation:

- *Work on participants' motivation, supplying them with key pieces of information regarding the study;*
- *Perform oversampling, i.e. sampling more than the minimum required;*
- *Plan to send reminders to the participants;*
- *Approach individuals personally when needed.*

KASUNIC's technical report (2005) presented a set of guidelines (hands-on) for conducting surveys in SE through a process composed of seven sequential steps, as already presented in Section 2.2. Although the technical report was published by the Software Engineering Institute, its content is predominantly addressed to introduce survey research in general. In this sense, similar to KITCHENHAM and PFLEEGER (2001, 2002, 2008), the researcher presents general principles regarding sampling activities, such as sampling methods and sample size formulas. On the other hand, the planning of subject recruitment is not discussed.

In order to investigate the participation of developers in SE surveys, SMITH *et al.* (2013) introduced a set of *persuasive factors* presented in Table 2-2. Such factors were borrowed from persuasive research (*reciprocity*, *consistency*, *authority and credibility*, *liking* and *scarcity*) and from recommendations observed in the survey literature in general (*brevity*, *social benefit*, *compensation value and likelihood* and *timing*). Using most of these factors would affect how the invitation message will be composed and sent. The researchers analyzed to which extent such factors were applied in a set of ten surveys having developers from Microsoft Company as subjects, observing that sending direct e-mail invitations (without using BCC) may influence on the increase/decrease of the response rate. Despite the small sample of surveys analyzed, the authors concluded that the presented factors could serve as a starting point for future studies on improving the response rates of SE surveys.

In addition to the previous methodological works introducing/ discussing the survey method, technical literature also presents a few experience reports discussing survey planning issues. Some of them consider the survey questionnaire composition (PUNTER *et al.,* 2002; CIOLKOWSKI *et al.,* 2003), while others discuss sampling issues (CONRADI *et al.,* 2005; JI *et al.,* 2008). CONRADI *et al.,* (2005) report in depth how the sampling frame for an international large-scale survey had been established through an

exhaustive process of gathering organizations' data from three countries (Germany, Italy and Norway) and using different data sources for each one, including Yellow Pages. Due to the limitation of information available in the used data sources, different ways for composing the survey sampling frame were applied. For instance, researchers called each organization listed in the Yellow Pages to identify which of them were active and working with the research theme. JI *et al.* (2008) replicated this survey in a fourth country (China), where a fourth different approach for sampling was applied. The authors then emphasized challenges to establish representative samples for SE surveys but recommendations to overcome such challenges were not provided.

Table 2-2. Persuasive factors that can influence the participation in surveys. Adapted from (SMITH *et al.,* 2013).

Description	**Trustworthiness attribute**
Reciprocity	Tendency on people complying with a request if they feel they owe the requester a favor
Consistency	Tendency on people continues to comply with agreements they have previously committed to. Consistency pressure can be used in surveys when we ask individuals if they would be willing to take a survey at some point in the future
Authority and Credibility	Compliance rates rise with the authority and credibility of the persuader
Liking	People are more likely to comply with a request from a person they have positive affect towards
Scarcity	It can be applied in survey research through establishing deadlines to participate in the study
Brevity	Long questionnaires should be avoided
Social Benefit	Potential participants may be more likely to respond to surveys if they see that their responses will benefit society, rather than a private entity
Compensation and Likelihood	Respondents may be promised compensation or the possibility of winning compensation, monetary or not
Timing	Is related with when (weekday, time) the recruitment activities are performed

Regarding survey reporting, STAVRU (2014) recently introduced a set of criteria of *thoroughness* to evaluate survey reports that can also be useful to evaluate the thoroughness of survey plans. Among others, such criteria include the need of papers providing the description of the survey *target population (audience)*, *sampling frame*, *sampling method* and *sample size*. Then, one or more of the following *trustworthiness attributes* should be applied to evaluate each criterion: *neutrality*, *consistency*, *truth value* and *applicability,* as presented in Table 2-2. The author concluded that eight out of the nine surveys investigating the use of agile methods do not present sufficient

thoroughness and subsequently low trustworthiness. Thus, in order to improve the observed scenario, STAVRU proposed a set of recommendations, including that *"special provisions should be taken to increase the objectivity of surveys on agile method usage in order to ensure that their findings are not biased by the individuals or organizations conducting them".*

Table 2-3. Thoroughness items and trustworthiness attributes applied in the survey plans' evaluation (STAVRU, 2014).

Thoroughness item	**Description**	**Trustworthiness attribute**
Target Audience	The study specifies and thoroughly describes its target audience	Truth value
		Consistency
Participants attributes	The study specifies the attributes that will be used to restrict the composition of the survey sampling frame and to characterize each participant	Truth value
		Consistency
Sampling frame	The study specifies and thoroughly describes its sampling frame, including from which source it will be obtained	Applicability
		Consistency
Sampling design	The study specifies and thoroughly describes its sampling design	Applicability
		Consistency
Recruitment Process	The study describes how the survey will be mediated	Consistency
	The study specifies how long the survey will be available to respondents	Consistency
	The study specifies how the invitations will encourage response and prevent non–response	Consistency

Thus, one can see that the guidelines analyzed in this section suggest how far we are from answering the research questions introduced by this thesis. With a few exceptions, the content of the analyzed publications typically addresses to survey research in general. Since literature reviews investigating how a survey method has been applied in the field were not found, we conducted a structured review on sampling and recruitment activities in SE surveys (DE MELLO and TRAVASSOS, 2015). The main findings of this work are presented in Chapter 3.

2.4. Conclusion

This chapter introduced the survey process and the main statistical concepts related to survey research followed by our research. More details of such concepts are not presented here as they are part of the different versions of the conceptual framework documentation. This chapter also presented and discussed publications used as guidelines to conduct surveys in SE. Since this chapter was focused in the related work

that inspired our research, we intentionally did not include the guidelines to conduct surveys in SE (LINÅKER *et al.*, 2015) derived from the survey course at SERG. In such work, some issues regarding the samples' representativeness in SE surveys also presented in this thesis were introduced, even though the conceptual framework is not presented.

3 Building and Evolving the Conceptual Framework

This chapter summarizes the research steps conducted before the development of the second version of the conceptual framework.

3.1. Introduction

This chapter presents the execution of the first four steps of the research methodology followed (Section 1.6) which includes the building of the first version of the conceptual framework (v1). The great majority of the research activities conducted in such steps have been already published. Thus, we avoid detailing them here, focusing on presenting their main findings.

3.2. Conceptual Technology Characterization

Based on the observed limitations to identify representative samples to surveys in SE and inspired by our previous experience in conducting SLRs (DE MELLO, PEREIRA and TRAVASSOS, 2010; DE MELLO *et al.,* 2014b), we depicted the initial concepts presented in Table 3-1 (DE MELLO and TRAVASSOS, 2013b) to be complementary to the statistical concepts of *target audience, sampling frame* and *sampling design* already used in the survey literature (and described in section 2.2). Such new concepts take into account a scenario in which sampling frames, suitable for a specific research topic, may be systematically retrieved from the content available in a selected *source of recruitment* (Figure 3-1). Examples of potential large-scale sources of recruitment (and respective search units) in SE research are typically available in the Web. They may include *professional social networks (groups of interest/ members)*, *discussion groups (members)*, *software projects repositories (project teams)*, *directories of software organizations/ professionals (members)* and *crowdsourcing tools (workers).*

Table 3-1. New concepts proposed to support the planning of surveys in SE.

Concept	Description
Target Audience	(see Section 2.2)
Source of Recruitment	Consists on a database (automated or not) from which valid subpopulations of the target audience can be systematically retrieved and randomly sampled
Search Unit	Characterizes how one or more survey units can be retrieved from a specific source of recruitment.
Technical Terms	Consist on a set of keywords connected through logical operators that can be applied to the source of recruitment in order to retrieve adequate search units
Search Criteria	Describe an algorithm to be followed in order to filter the search units in a source of recruitment, including how to apply the planned search string
Exclusion Criteria	Describe a set of restrictions that must be applied in order to exclude undesirable retrieved search units
Sampling Frame	(See Section 2.2)
Sampling Design	(See Section 2.2)
Recruitment Process	Describes how the survey sample will be obtained from the sampling frame and how subjects will be recruited

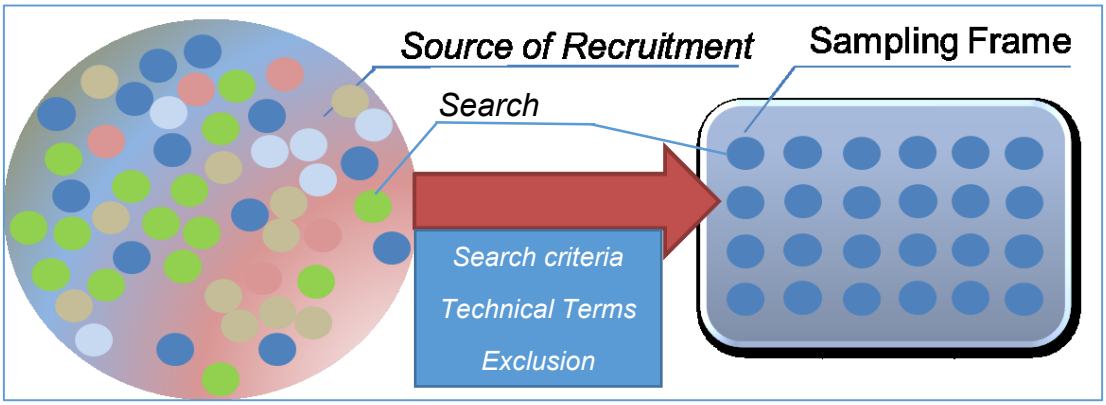


Figure 3-1. Representation of a sampling frame composition from a source of recruitment.

3.3. Preliminary Studies

The concepts presented in Section 3.2 were then applied to *recruitment plans* to support the re-execution of three distinct surveys planned by researchers from the Experimental Software Engineering group at COPPE/UFRJ. For each survey re-execution, it was expected that its recruitment plan could evolve the original survey design by describing the way more representative samples would be systematically obtained from generic sources of recruitment available in the Web.

Since samples' *heterogeneity* and the *confidence level* of subjects' responses are key factors to evaluate samples' representativeness, the following null and alternative

hypotheses were established to support the comparison between the set of respondents (*effective sample*) obtained in the previous survey execution(s) (ES1) and the set of respondents obtained in the new execution following a recruitment plan (ES2):

- $H_0$1: *ES1 and ES2 present equal heterogeneity;*
- H_A1: *ES2 is more heterogeneous than ES1;*
- $H_0$2: *ES1 and ES2 present equal confidence level;*
- H_A2: *The confidence level of ES2 is higher than ES1.*

The following subsections summarize each preliminary study, presenting its recruitment plan and discussing its main findings.

3.3.1. Survey on Requirements Effort Influence Factors

The first preliminary study was conducted in the context of a survey on requirement effort influence factors (VAZ, 2013). The main goal of this survey is to identify factors that can influence the effort involved to conduct requirement activities. For this study, the *heterogeneity* of each effective sample was analyzed based on the following properties:

- The *geographical distribution of the sample* (country);
- The diversity of *most experienced domains* of the survey respondents.
- The distribution of *software requirement techniques and approaches* followed by the respondents.

The *confidence level* of the contribution from each respondent was calculated based on his/her *academic degree* and his/her *experience with requirement engineering activities,* applying the following formula (1):

$$F(n) = \frac{\sum_{n=1}^{6} L(An) + L(Acad) + L(Nproj)}{Max} \quad (1)$$

- L(An): *experience level answered for each one from six (An) requirement engineering activities, such as requirement elicitation with various stakeholders and working with various requirement engineering teams. (Likert scale: 0- None, 1-Low 2-Medium, 3-High)*
- L(Acad): *higher academic degree (scale: 0- high school, 1-undergraduation 2-specialization, 3-Msc. and 4-D.Sc./ Ph.D.);*

- L(NProj): *estimated number of projects that the subject participated (scale: 0, if none; 1, 2, 3 or 4 based on the quartiles for the whole distribution, including ES1 and ES2).*
- Max: *represents the maximum value of the numerator.*

While the original sample (S1) was undetermined - it was established by convenience through sending invitation messages to researchers' personal/ professional contacts and allowing redirection to other individuals - S2 had followed the recruitment plan presented in Table 3-2. After executing the recruitment plan, a set of search criteria was applied and 14 groups concerned with requirements engineering were randomly selected out of all retrieved groups. Initially, the originally planned recruitment, which regards the posting of generic invitations in group forums was followed. However, after one week, only 18 subjects had answered it. In addition, we concluded that such recruitment approach hampers the control of the sample representativeness, being, then, discarded. Thus, we decided to change the recruitment plan *on the fly* by manually sending individual invitation messages to 924 distinct members from two randomly selected groups of interest. After two weeks of the survey execution, an effective sample (ES) composed of 34 subjects was obtained, totalizing 52 participants in ES2 against 32 participants in ES1. A comprehensive description of this preliminary study and discussion of its results can be found in (DE MELLO and TRAVASSOS, 2013b).

Table 3-2. Recruitment Plan designed for S2. Based on (DE MELLO and TRAVASSOS, 2013)

Concept	**Description**
Target Audience	Professionals working with software requirements
Source of Recruitment	LinkedIn
Search Unit	Groups of interest created by LinkedIn members in which other members can subscribe, becoming able to participate in forums and perform another social activities.
Technical Terms	"Requirements Traceability", "Requirements Engineering", "Requirements Verification"; "Requirements Validation", "Requirements Elicitation"; "Requirements Specification", "Requirements Management"; "Software Requirements"; "Requirements Analysis", "Requirements Review", "Functional Requirements", "Non- Functional Requirements", "System Requirements", "Requirements Inspection", "Requirements Gathering", "Domain Requirements"
Search Criteria	For each technical term, submit the term (between quotes) using the "Group Search" option. Then, select the first and second groups listed in the results.
Exclusion Criteria	After executing the search criteria, it shall be analyzed each suggested and selected group's name, description and its members. To ensure

	the necessary access for recruitment and to avoid authors' influence bias we have decided to exclude the groups that: – Do not accept at least one of the survey researchers as member until the survey execution date; – Explicitly prohibit the use of the forum for the recruitment of subjects for experiments; – Have more than 10% of its members directly connected with the researchers' profiles, and; Have the active participation of the researchers over the last month before execution
Sampling Frame	The sampling frame will be composed by the members from all groups of interest selected.
Sampling Design	Census, since each member from each group of interest selected is able to participate in the survey
Recruitment Process	For each selected and included group of interest, it shall be posted a recruitment message on its forum based on a specific template.

3.3.1.1. Heterogeneity Analysis

Regarding the geographical distribution, it was possible to observe that S2 subjects come from more diverse countries (17) than S1 (8). Table 3-3 presents a comparison between the distributions of both samples by geographical regions. Since Brazilian researchers conducted the survey, one can see how the geographical distribution of their personal contacts had concentrated ES1 in Latin America. In the other hand, S2 suggests a more realistic and appropriate geographic distribution to support observations.

Table 3-3. Distribution of each effective sample among geographical regions. (DE MELLO and TRAVASSOS, 2013b)

Geographical Region	ES1	ES2
Latin America	81.25%	9.62%
Europe	12.50%	44.23%
USA and Canada	3.13%	36.54%
Oceania	3.13%	0.00%
Asia	0.00%	9.62%
Africa	0.00%	0.00%

Regarding more experienced domains reported by subjects, we identified that the ones mentioned the most from both samples are concerned with BFSI- Banking, Financial Services and Insurance (32.14% from S1; 27.59% from S2). However, it is also possible to observe that S1 presents a higher concentration of subjects reporting more experience on e-commerce (14.29%) and information systems (14.29%) domains, while S2 presents a higher concentration of experienced subjects on military (18.42%) and government (13.16%) domains. The experience of S2 subjects in 13 distinct requirement

techniques and approaches is similar to the distribution of most frequent approaches (*ad-hoc*, RUP-based and agile-based) in S1.

Thus, although we did not observe great differences in technologies and domain experiences between both samples, the observed difference regarding geographic dispersion between ES1 and ES2 can support our decision to refuse $H_0 1$ and accept $H_A 1$. Therefore, ES2 is more heterogeneous than ES1.

3.3.1.2. Confidence Analysis

The subjects were asked about their experience level (None-Low-Medium-High) regarding the requirement engineering activities presented in Table 3-4, as well as their higher academic degree. We applied the *chi-square* test for comparing the frequencies of the experience level between ES1 and ES2 for each activity answered. Due to the low frequencies observed in the lower levels (None, Low), we decided to group them in just one level to make the tests feasible. *Chi-square* tests revealed that ES2 frequencies are significantly different from ES1, with *p-values* less than 0.003 for A1, A3, A4, A5 and A6, and *p-value* = 0.037 for A2. While the median for S2 in all activities was "High", S1 median was "None-Low" for A1 and A4, and "Medium" for A2, A3, A5 and A6. When applying the same test for "higher academic degree" it wasn't observed any significant difference between S1 and S2.

Table 3-4. Results of the Chi-Square test for each attribute (ES1 X ES2).

Attribute	**Chi-Square test**	**Degrees of freedom**	***p*-value**
Requirements effort estimation	26.651	2	< 0.001
Requirements specification in different problem domains	6.572	2	0.037
Requirements engineering techniques and processes	13.598	2	0.001
Requirements project management	14.185	2	0.001
Requirements elicitation with various stakeholders	13.292	2	0.001
Working with various requirements engineering teams	11.865	2	0.003
Higher Academic Degree	4.873	4	0.301

Figure 3-2 shows the distribution of the calculated experience level for both individual samples ES1 and ES2, and for a combined sample ES1 + ES2, based on the calculated experience rate. We tested distribution normality (Kolmogorov-Smirnov) and

observed that ES2 distribution is not normal with *p-value*< 0.01 while ES1 distribution is normal. Applying the non-parametric Mann-Whitney test (two-tailed), we observed that ES2 has more experienced subjects than ES1. The test is significant at 0.0001. Thus, we are able to reject $H_0$2 and accept H_A2. Therefore, ES2 confidence level is higher than ES1.

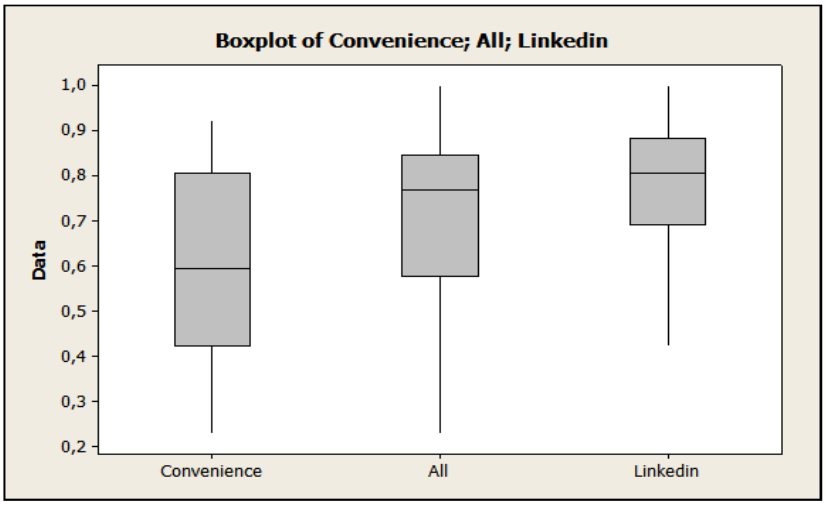


Figure 3-2. Distribution of experience level for ES1 (convenience), Both (ES1 + ES2) and ES2 (Linkedin).

3.3.1.3. Discussion

Although the professional social network LinkedIn was not designed to support survey research, we observed that its use was helpful to reduce bias on sampling activities. However, the recruitment performed through sending generic invitations in the groups' forums is prone to be ineffective in providing representative samples, since there is no control on who reads the recruitment message. In addition, we observed that relatively few members effectively participate in such forums. Thus, we can infer that the results following such approach could be strongly influenced by the subjects' effective participation in the social network.

On the other hand, sending individual messages in large scale was helpful to increase the sample representativeness, despite the unexpected small participation rate. However, manually performing such activity may bring undesirable operational risks. In this sense, providing automatized support to operate individual recruitment activities in LinkedIn could mitigate them. Regarding the subjects' characterization, one can suggest LinkedIn members tend to overestimate their skills; however, the experiences of the ESE Group when conducting previous surveys over convenience samples (personal contacts) have shown that overestimation bias also happens in this context.

In general, the survey re-execution was helpful to strengthen evidence regarding the survey topic. However, few but significant differences of opinion were identified between ES1 and ES2. After analyzing such differences, we observed they could be influenced by their significant difference of experience in requirements engineering activities.

3.3.2. Evaluation of Guidelines for Simulation-Based Studies in SE

The second preliminary study was performed in the context of the evaluation of guidelines for simulation-based studies in SE (FRANÇA and TRAVASSOS, 2015). In this empirical study, subjects were invited to first read a document describing such guidelines and, then, answer a survey questionnaire. In this context, the evaluation of respondents' heterogeneity was done considering their geographical distribution, while each respondent's confidence level was calculated based on their reported higher education level and on their reported experience with SE research and simulation studies. S1 was composed by two distinct sources: authors of relevant papers identified through a SLR considering the research theme and a set of researchers sampled by judgment. On the other hand, S2 was composed of the execution of the recruitment plan presented in Table 3-5.

Table 3-5. Recruitment Plan designed to S2.

Concept	Description
Target Audience	Software Engineering Researchers working with simulation
Source of Recruitment	The professional social network ResearchGate[7]
Search Unit	ResearchGate members
Technical Terms	"Agent-Based Simulation", "Conditional Growth", "Discrete-Event", "Event-Based Simulation", "General Continuous-time Simulation", "General Discrete-time Simulation", "Hybrid Simulation", "Monte Carlo", "Object-Oriented Simulation", "Proxel-based Simulation", "Qualitative Simulation", "quasi-Continuous Simulation", "Semi-Quantitative Simulation", "State-based Simulation", "Stochastic Process Algebra", "System Dynamics", "Temporal Parallel Automata", "Computer Simulation", "in silico", "in virtuo", "modeling and simulation", "Knowledge-based simulation", "sampling", "simulation and modeling", "stochastic modeling", "simulation study"
Search Criteria	For each technical term do: Submit the term between quotes in the "Search" option, combined with the expression. "Software Engineering" (between quotes)

[7] www.researchgate.net

	Select all members retrieved by the search.
Exclusion Criteria	None
Sampling Frame	The list of all ResearchGate members selected and included.
Sampling Design	Simple Random Sampling, confidence level of 95% and confidence interval of 3 points
Recruitment Process	It will be sent individual invitation messages through the ResearchGate message service.

As a result, a sampling frame composed of 1,000 ResearchGate members was identified. Out of those, 300 were randomly sampled and recruited. However, although the survey target audience was composed of researchers, the participation rates observed in both samples were significantly low (only two subjects in ES1 and 11 subjects in ES2). One possible reason for such low participation rates is addressed to the fact that the respondents had to read a 22-page document before answering the survey questionnaire. As a consequence, we were not able to test the preliminary study hypotheses. Nonetheless, ES2 subjects provided relevant contributions to improving the simulation guidelines (FRANÇA and TRAVASSOS, 2015).

The experience obtained in this study allowed us to observe the need to investigate better the source of recruitment selected before using its resources in the survey execution. The recruitment activities were limited by unexpected ResearchGate restrictions on sending individual messages in large-scale, which led to operational errors (we identified that the same subject was recruited in both S1 and S2). Notwithstanding, it is important to point out that ResearchGate was useful to identify a relevant set of SE researchers taking into account the specificity of the research theme. In this sense, it is important to point out the relevance of using a suitable set of technical terms.

3.3.3. Survey on Agile Characteristics and Agile Practices in Software Processes

A third recruitment plan was depicted to support the re-execution of the survey designed to evaluate the pertinence and relevance of characteristics of agility and agile practices to introduce agility in software processes (ABRANTES and TRAVASSOS, 2013). Due to the restrictions imposed by the original survey questionnaire (elaborated in 2011), the sample heterogeneity was evaluated only using the *geographical distribution* of the samples while the *confidence level* of each subject was calculated using the following formula (2):

Exp= (L(Acad)+L(AgPap)+L(AgExp)+S(AgProj))/Max (2), where:

- Acad: *Academic Degree, (Likert scale: 0- Undergraduation, 1-Specialization 2-Master Degree, 3-DSc/PhD Degree);*
- AgPap: *number of publications related to agile context, (scale: 0-3, based on the quartiles distribution);*
- AgExp: *perceived experience on agility in software process, (Likert Scale: 0-Low, 1-Medium 2-High, 3-Very High);*
- AgProj: *amount of Agile Projects, (scale: 0-3, based on the quartiles distribution);*
- Max: max value of the *numerator.*

For this preliminary study, S1 was composed of 158 authors of papers identified on a SLR conducted to organize an initial set of characteristics and agile practices and of the original signatories of the agile manifesto (ABRANTES and TRAVASSOS 2013). Alternatively, the recruitment plan synthesized in Table 3-6 was followed in order to compose S2, resulting in a sampling frame composed of more than 150,000 members from 19 LinkedIn groups of interest. Taking into account the similarities observed among such groups (the members overlapping, i.e., the distributions of members in common), they were organized into eight strata resulting in 7,745 subjects recruited.

Table 3-6. Recruitment Plan designed for S2. Based on (DE MELLO, DA SILVA and TRAVASSOS, 2014)

Item	**Description**
Target Audience	Software Engineering practitioners involved with agile
Source of Recruitment	LinkedIn
Search Unit	Group of Interest
Technical Terms	*“agile”, “agility”, “test-driven development”, “continuous integration”, “pair programming”, “planning game”, “on site customer”, “Collective Code Ownership”, “Collective ownership”, “small releases”, “short release”, “developing by feature”, “metaphor”, “refactoring”, “Sustainable Pace”, “simple design”, “coding standards”, “whole team”, “project visibility”, “daily meetings”, “open workspace”, “product backlog”, “planning game”*
Search Criteria	For each keyword, do: 1. Submit a search expression (between quotes) followed by the term “software” in the option “Group Search”; 2. Identify all groups of interest returned, recovering the following data: name, description, group rules and number of members.
Exclusion Criteria	It will be excluded groups that: – explicitly prohibits the execution of studies; – explicitly restricts the individual messaging between its members (a default feature provided by LinkedIn);

	– explicitly directs to a city, region or country, since our target audience are not geographically restricted; – is focused on promoting specific organizations, or provided by them, neither to disseminate specific events; – has its description out of the scope of Software Engineering; – has a vague description; – has a single member; – is driven to headhunting and job offering; – represents LinkedIn' subgroups, since the sampling frame must be composed by groups of interest, and; – has a non-English language as default, since English language is default in international forums.
Sampling Frame	The list of members from the more representative groups found in the search
Sampling Design	Based on the similarity observed between groups, to evaluate the feasibility of clustered sampling. If not possible, to apply stratified sampling
Recruitment Strategy	To send individual invitations based on a default recruitment message using the resource of sending messages for group members provided by LinkedIn

As a result, ES2 was composed of 291 respondents. Based on the main SE skills reported by them in the survey questionnaire, the respondents were then re-organized into the following five strata (DE MELLO, DA SILVA AND TRAVASSOS, 2014b):

- *Agilists*, composed of many *LinkedIn* groups concerned with agility in SE. The main skill groups are: *personal skills* (11.06%), *social skills* (10.38%) and *Software analysis and design* (9.10%);
- *Testing Professionals,* mainly composed by *LinkedIn* groups devoted to *Software Testing*, also representing the most relevant skill group (14.80%);
- *Programmers,* mainly composed of *LinkedIn* groups devoted to *agile practices*, having *programming* as the most relevant skill group (15.76%);
- *Configuration Managers*, composed of three *LinkedIn* groups concerned with configuration management (CM). The main skill groups are: *CM* (12.30%), *programming* (10.73%) and *personal skills* (10.09%),and;
- *System Analysts*, composed of a single *LinkedIn* group devoted to *software architecture*. The main skill groups are: *personal skills* (15.53%) and *Software analysis and design* (14.26%).

A comprehensive description of this preliminary study can be found in (DE MELLO, DA SILVA and TRAVASSOS, 2015). However, other publications present and discuss more in depth the recruitment plan execution (DE MELLO, DA SILVA and

TRAVASSOS, 2014) and the preliminary study results (DE MELLO, DA SILVA and TRAVASSOS, 2014b).

3.3.3.1. Heterogeneity Analysis

ES2 sample was composed of respondents from 43 distinct countries, distributed in all geographic regions (Table 3-7) while respondents from only nine distinct countries were found for ES1. Although approximately half of ES1 subjects (13) did not provide their location (it was an open question), one can see that even if all ES1 subjects (25) had reported a distinct country, it could not be possible to overcome the heterogeneity observed in ES2. Thus, $H_0 1$ cannot be accepted and $H_A 1$ becomes valid.

Table 3-7. Distribution of ES1 and ES2 by Geographic Region.

Geographic Region	**ES1 (12 respondents)**	**ES2 (289 respondents)**
North America	41.7%	38.1%
Europe	33.3%	41.2%
Asia	25%	11.8%
Latin America	-	5.9%
Oceania	16.7%	2.1%
Africa	-	0.1%

Additionally, we had observed that the distribution of members from the 10 most represented countries in the population from which S1 were obtained presents a high correlation (95.1%, p-value <0.001) with the distribution of participants from the same countries in ES2 (Pearson's correlation test). These findings suggest that the participation in the survey was not strongly influenced by language barriers[8] since five of these 10 countries do not have English as a native language. Although we also evidenced the heterogeneity of ES2 through the diversity of the main SE skills reported, we were not able to compare these results since such information was not collected in ES1.

[8]However, we cannot infer to which extent language barriers influence subscription in *LinkedIn* and in its groups of interest.

3.3.3.2. Confidence Analysis

After calculating the confidence levels for ES1 and to all ES2 strata, we observed that ES1 do not follow a normal distribution (Shapiro-Wilk)[9]. Thus, we applied the non-parametric Mann-Whitney compare the distributions of confidence level (Table 3-8). The number of outliers removed from each distribution is indicated in parenthesis.

Table 3-8. Comparison of Experience Level distributions.

Sample	Sample Size	Mean	Median	Mann-Whitney test (greater than ES1?)
ES1	25(1)	0.375	0.333	-
ES2	291(13)	0.416	0.417	0.027
Agilists	97(0)	0.460	0.500	0.007
Testing professionals	81(3)	0.382	0.417	0.170
Programmers	57(7)	0.470	0.500	0.001
Configuration Managers	24(0)	0.351	0.333	0.415
System Analysts	31(0)	0.403	0.417	0.168

Hence, we observed that ES2 distribution is significantly greater than ES1, although the same behavior could be observed in only two strata: *agilists* and *programmers*. We also compared the confidence level distributions of these two strata with the other ones. As a result, we noticed that both agilists and programmers are significantly more experienced than testing professionals, configuration managers and system analysts. The presented results led us to reject $H_0 2$ and to accept $H_A 1$.

3.3.3.3. Discussion

Different from the previous preliminary studies, ES2 subjects were recruited using automated support (macros) which helped us mitigate the operational risks involved on sending individual messages manually. Regarding the sampling design, it is important to highlight that stratification of groups of individuals sharing members in common is not recommended in stratified sampling. However, we understand that the intensity of such behavior in social networks could suggest more or less similarity among them. In this sense, we controlled the sampling activities in such a way to avoid sampling the same

[9]In the original analysis we wrongly classified distribution of ES1 as normal and homoscedastic, which led us to perform some comparisons applying the t-Student test. Although the same main results regarding the comparison of ES1 and ES2 strata were obtained, Mann-Whitney test revealed that distribution of the whole ES2 confidence level is significantly greater than ES1.

subject twice in different strata. The investigation performed in the third preliminary study allowed us to learn some important lessons regarding sampling in SE surveys, such as the following (DE MELLO, DA SILVA and TRAVASSOS, 2014b):

- Since social network groups of interest are naturally established, each group can initially be considered a stratum. Then, the similarities observed can be used to group them;
- Although the overlapping of members can suggest similarity among groups, it is risky to rely only in such property to perform the stratification;
- It is worthwhile investigating in depth the characteristics of groups of interest available in professional social networks;
- A simple and optional open question can make a significant difference and contribute to trace subjects' profiles without overloading them with many characterization questions. However, the coding process in large scale can be exhaustive and it must be considered in the study planning. In addition, such coding is biased by the interpretation of the researchers.

In addition, bearing in mind the findings of the three preliminary studies conducted, we observed that voluntary participation of random samples in surveys tends to be small in SE. However, a small participation rate on a large scale can be better than a large rate on a small scale.

As to contributions to the survey topic, the diversity of ES2 samples allowed us to observe that introducing agility into software processes is more complex than an initial interpretation could suggest. We could observe that researchers (ES1) and practitioners (ES2) reached some consensus when it comes to indicating those characteristics of agility that can contribute more or less to the introduction of agility into software processes. In the case of the agile practices, their opinion diverged more, which can be explained by the significant difference of practical experience observed among samples. The stratification performed in ES2 helped us identify how the respondents' background could influence their opinion. For instance, the *agilists'* stratum has been the group that agreed the most with the high relevance characteristics of agility and agile practices evaluated; whereas *testing professional*s was the stratum that better evaluated *continuous testing* characteristic and *test-driven development* practice. Thus, we emphasize the relevance of investigating to which extent the interpretation of "agility" by an individual could be based on his/her strictly personal interpretation, limited to their software engineering abilities or whether it is also based on their conscious and holistic observation of the software processes and the state of practice. A comprehensive

discussion about the survey results can be found in (DE MELLO et al., 2014e; DE MELLO and TRAVASSOS, 2016b).

3.3.4. Threats to validity

Although each preliminary study has its own threats to validity (discussed in different publications), it is important to point out some general issues. First, one can suggest that it was invested more effort to re-execute the surveys than the original executions, which could be related with the (in general) better re-executions results. However, such effort was more concentrated in activities concerned with generic survey practice than in designing the recruitment plan. For instance, great part of the research effort in the three studies was spent in the execution of manual recruitment activities, since the used sources of recruitment were not ready to support the automated sending of invitations. In addition, the third study had demanded great effort also due to the complex sampling analysis performed, involving sampling designs not commonly used in the SE field (clustered/ stratified sampling).

Second, once the three studies were grounded in surveys already planned/executed, the comparison among the subjects' characteristics from different executions was limited to the items included in the original survey questionnaires. However, we understand that subjects' location (country, used in the three surveys) can be a good indicative of heterogeneity in the context of a socio-technical field such as SE mainly due to the cultural aspects that can influence the subjects' opinion regarding different research topics.

The third issue is concerned with the limited control of the survey population available due to the dynamic and restrictive characteristics of searching and retrieving data in the used sources of recruitment, especially in the professional social network LinkedIn. Once we did not have access to the whole population members, we worked with the subsets of samples retrieved by the professional social network for each selected group of interest. In (DE MELLO, DA SILVA AND TRAVASSOS, 2015) we describe how we mitigated such bias through analyzing the personal connections over the subsets of samples retrieved by LinkedIn.

3.4. Conceptual Framework- First Version

Based on the lessons learned from applying recruitment plans in the preliminary studies, we designed the first version of the conceptual framework (v1). In order to better organize the knowledge, some concepts presented in Section 3.2 were reviewed, new concepts (and properties) were added and relationships among them were established.

Table 3-9 presents the conceptual framework v1 concepts, pointing out such differences. A detailed description of the conceptual framework v1 can be found in (DE MELLO et al., 2014).

Table 3-9. Conceptual Framework v1 concepts and properties.

Conceptual Characterization of the Technology	**Conceptual Framework v1**	
	Concept	**Properties**
Target Audience	Target Audience	-
-	Unit of Observation	Unit of Observation Attribute
Source of Recruitment	Source of Sampling	-
Search Unit	Search Unit	-
Technical Terms	Search Plan	Search String
Search Criteria		Search Algorithm
Exclusion Criteria		Exclusion Criteria
Sampling Frame	Sampling Frame	-
Sampling Design	Sampling Strategy	Sampling Design, Confidence Level and Confidence Interval
-	Measurement Instrument	-

The statistical concept of *unit of observation* (WOHLIN ET AL., 2012) was included in the conceptual framework to clearly restrict who is able to participate in the survey and to establish which attributes should be used to characterize each subject. If one or more of such attributes are not available in the source of sampling, a specific *measurement instrument* (WOHLIN ET AL., 2012) should be provided in order to retrieve such data. In the context of survey research, the measurement instrument is typically an initial section in the survey questionnaire.

To support the evaluation of candidates to the sources of sampling and their respective search units, *essential requirements* and *desirable requirements* were designed. Such requirements were, then, used to evaluate different candidates to the sources of sampling available in the Web, including professional *social networks*, *crowdsourcing tools* and *freelancing tools* (DE MELLO et al., 2014). Such evaluation indicated the unfeasibility to use the analyzed crowdsourcing tools (Mechanical Turk, Micro Workers and Crowd Workers) mainly due to the fact that such tools do not allow the sampling process to be controlled, providing only blind recruitment. A broader discussion regarding the limitations to use MTurk to support large-scale studies is presented in (DE MELLO, STOLEE and TRAVASSOS, 2015). Professional social networks may be used as a source of sampling but being mindful of the technical limitations imposed by each one for searching units and/or recruiting subjects, as already

exemplified in Section 3.3. Finally, it is important to point out that freelancing tools could provide a rich environment to support sampling and recruiting activities, but it may also be expensive to hire professionals in large scale using such tools to answer surveys.

3.4.1. Proof of Concept

As already mentioned in the Introduction of this Thesis, the sampling issues studied are not exclusive from survey research, also affecting large-scale experiments. Thus, the conceptual framework v1 was applied to support the replication of an experiment on Java code search sampling activities (STOLEE, ELBAUM and DWYER, 2015). Before participating, subjects should be invited to perform a qualification exam on Java programming. Then, only the qualified subjects could participate. Such experiment was replicated twice but only in the first trial, subjects had answered the characterization questionnaire presented Table 3-10 before executing the experimental task.

Thus, we planned to compare the characteristics of the effective sample obtained in the first trial (ES1), not using the conceptual framework, to the effective sample obtained in a third trial (ES2), using the conceptual framework. We also planned to compare the performance of the candidates from both trials in the qualification exam. Thus, the following hypothesis emerged:

- *$H_0$1: There is no association between the source of the potential subjects and the qualification exam results*
- *H_A1. There is an association between the source of the potential subjects and the qualification exam results*
- *$H_0$2. There is no difference between the experience level of ES1 and ES2*
- *H_A2. The experience level of ES2 is different from ES1*
- *$H_0$3. There is no difference between the programming habits of ES1 and ES2*
- *H_A3. ES2 has different programming habits from ES1*

Table 3-10. The Subject Characterization Questionnaire (DE MELLO, STOLEE and TRAVASSOS, 2015).

Q1: How many years of programming experience do you have?
Q2: How many years of Java programming experience do you have?
Q3: How often do you program?
( ) daily
( ) weekly
( ) monthly
( ) never
Q4: How many search results do you typically examine before finding something useful?
Q5: How many different search queries do you try before finding a useful result?

ES1 was obtained by convenience, composed of 19 self-selected Amazon's Mechanical Turk workers. Consequently, no sampling frame was established. On the other hand, when following the instantiation of the conceptual framework presented in Table 3-11, a sampling frame composed of 165,134 professionals spread over 40 countries from all continents could be obtained. From this, 1,647 subjects were recruited and 83 effectively participated, composing ES2.

While ES1 subjects were paid by each experimental task performed, donations to Brazilian Red Cross were made in the name of those subjects from ES2 that performed all eight experimental tasks assigned to them. The complete description of the presented study and a comprehensive comparison between the protocols and the results of the experiment trials is presented in (DE MELLO, STOLEE and TRAVASSOS, 2015).

Table 3-11. The instantiation of the conceptual framework v1 to support the experiment replication.

Concept	**Property**	**Description**
Target Audience	-	Java programmers
Unit of Observation	Unit of Observation Attributes	– Location (country) – Programming Experience (years) – Java Programming Experience (year) – Programming frequency – Searching code habits
Source of Sampling	-	LinkedIn
Search Unit	-	Group of Interest
Search Plan	Search String	"Java programming"
	Search Algorithm	Submit the search string (between quotes) in the option "Group Search"; Select the biggest group (in number of members) available
	Exclusion Criteria	none
Sampling Frame	-	The list of members from the selected group working with Information Technology, Computer Software and Telecommunications.
Sampling Strategy	Sampling Design	Stratified Sampling, by subject location: *Asia, USA+Canada, Europe, Latin America, Africa, Oceania*
	Confidence Level	95%
	Confidence Interval	0.06
Measurement Instrument	-	Java Qualification Exam + Subject characterization questionnaire from the original survey execution

3.4.1.1. Qualification Exam Analysis

Less than 3% of the candidates from ES2 were not approved in the exam, while more than 10% of the subjects from ES1 were. When applying the Pearson's chi-square test over the distributions, it was found that the performance in the qualification exam is associated to the source of the candidate, with a p-value of 0.018. Thus, it was possible to refute $H_0 1$ and accept $H_A 1$.

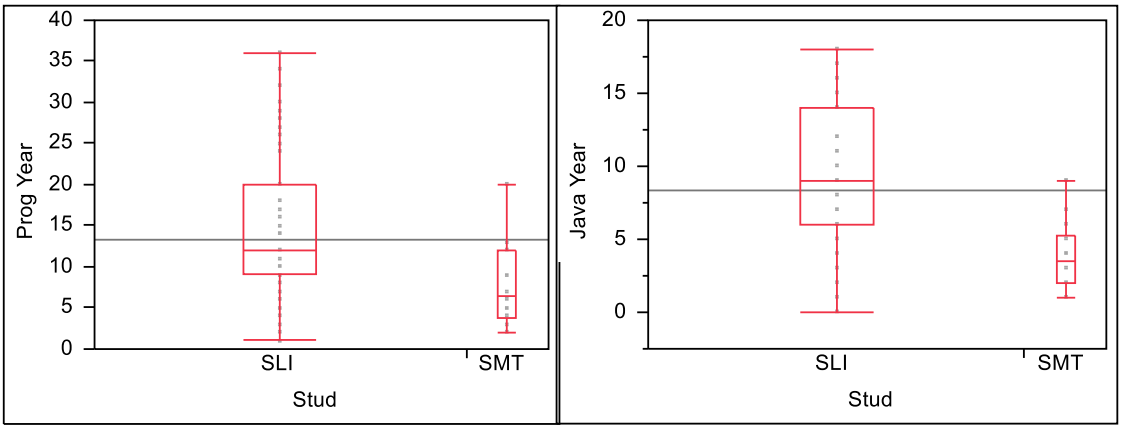
3.4.1.2. Experience Level Analysis

After removing three outliers, the ranges, the medians and the means for the distributions of programming experience were calculated, as presented in Table 3-12 (outliers removed in parenthesis). Figure 3-3 (a) presents the distribution of the programming experience in both samples. One can see that the range of years of programming in ES1 is contained in the larger range of years in ES2, suggesting that the distribution in ES1 is more diverse. The results from the Mann-Whitney test indicate that ES2 has a significantly higher distribution of programming experience than ES1 with a p-value of 0.0004.

Table 3-12. Programming Experience by Sample

Sample	Size	Mean	Standard Deviation	Median	Minimum	Maximum
ES1	18(1)	7.39	4.79	6.50	2	20
ES2	81(2)	14.62	8.41	12.00	1	36

Regarding Java programming experience, just a single outlier was removed. As Table 3-13 shows, the distribution of ranges suggests that ES2 is more diverse than ES1. At the same time, the boxplots presented in Figure 3-3(b) show how ES1 is concentrated in the range of 1-4 years of Java programming experience. The results from the Mann-Whitney test indicate that ES2 has a significantly higher distribution of Java programming experience than ES1 with p-value< 0.0001. Thus, the results observed on the experience level allowed us to reject $H_0 2$ and to accept $H_A 2$.



(a) (b)

Figure 3-3. Programming experience and Java programming experience by sample.

Table 3-13. Java programming experience by sample.

Sample	Size	Mean	Standard Deviation	Median	Minimum	Maximum
ES1	18(1)	4.06	2.46	3.50	1	9
ES2	83(0)	9.28	4.99	9.00	0	18

3.4.1.3. Programming Habits

Most of respondents in ES1 and ES2 have the habit of programming *daily* (68% and 70%, respectively). Since *weekly* (ES1), *monthly* (ES1) and *never* (ES1 and ES2) distributions values present insufficient sizes to apply the chi-Square test, we combined such values in a single one (not daily) which reduced the degrees of freedom of the test to two. As a result, evidence regarding the influence of the samples in the frequency of programming reported by the subjects was not found.

With regards to the characterization question *"How many search results do you examine…"* (Table 3-14), the results from the Mann-Whitney test indicate that ES2 has a significantly different distribution of number of searches in ES1 with p-value= 0.0017. Concerning the question *"How many search queries do you try..."*, we observed that ES1 and ES2 present similar ranges and close means. Applying the Mann-Whitney test, no significant difference among the distributions was found (p-value=0.1168). Therefore, considering the results observed only about the number of search results, it was possible to reject $H_0 3$ and to accept $H_A 3$.

Table 3-14. Number of search results by sample.

Sample	Size	Mean	Standard Deviation	Median	Minimum	Maximum
SMT2	14(5)	2.429	0.64	2.5	1	3
SLI1	69(14)	3.696	1.53	3	0	8

3.4.1.4. Discussion

It is not possible to infer to which extent a result obtained from a random sample is representative if the population involved is unknown. Evaluating web-based sources of sampling helped us observe that it can be helpful to avoid limitations on controlling the sampling process. However, the decision of selecting a specific source of sampling can also be influenced by factors related to the costs involved and even to the stability of the content and resources available in the source. For instance, in 2014, our experience showed that by using a (paid) premium account on *LinkedIn* we could access more subjects (DE MELLO, DA SILVA and TRAVASSOS, 2015) than by using a basic account. However, it was recently observed that *LinkedIn* no longer allows members from a group (independently from its account type) to keep in touch with a large set of other members of the same group, restricting such possibility to 10 members. This change in the rules highlights the need for a careful evaluation of the source of sampling selected before using it.

Our experience applying the framework on an online experiment helped us to identify some gaps, especially the need to provide guidance for recruitment activities. In our opinion, donating was important for stimulating the participation in the study. In fact, the participation rate in this case (5.01%) was higher than those in the re-executed surveys (3.68% – 3.76%) using the same source of sampling (subsections 3.3.1, 3.3.2 and 3.3.3) and similar recruitment steps. It can be considered a relevant result, once experimental tasks are supposed to demand more effort than surveys.

3.5. Structured Review

In order to support the evolution of the conceptual framework providing guidance for applying its concepts, we conducted a structured review over EASE/ESEM proceedings (electronically available, 2005-2014) on how sampling activities have been conducted in SE surveys. By using SCOPUS[10] as search engine, it was retrieved 82

[10] www.scopus.com

papers published between 2005 and 2014 citing “survey” in the title, abstract and/or keyword. After reading each paper abstract, 54 papers effectively reporting (opinion) surveys were included. The full reading of these papers allowed us identifying 52 distinct surveys. From these, only in 39 papers (37 surveys) we could distinguish the survey unit of analysis from the survey search unit.

Table 3-14 presents the distribution of these 39 papers classified by arrangements of unit of analysis- search unit type, distinguishing papers that reported the use of probabilistic sampling designs from those reporting non-probabilistic sampling designs. The complete description of this review can be found in (DE MELLO and TRAVASSOS, 2015).

In general, it was observed that most of surveys analyzed had been conducted using samples of *individuals* identified by convenience as unit of analysis. Although 19 from the 39 papers analyzed reports surveys using *individuals* as both search units and units of analysis, we identified only three surveys applying some probabilistic sampling design over a sampling frame formally established. BEGEL and NAGAPPAN (2008) randomly recruited 2,821 individuals having the list developers from a single organization (Microsoft) as sampling frame, approximately 10% from the total population available. MURPHY *et al.* (2013) performed another survey in the same organization, collecting the impression of software developers regarding agile practices at the company during six years. The work presented by RODRÍGUEZ *et al.* (2012) exemplifies the benefits to our field when a nationwide database is available. Through using data from the Finnish Information Process Association (FIPA) as source of sampling, researchers established a sampling frame composed by 4,450 Finnish SE practitioners suited to the survey focus.

We also identified the use of diverse alternative sources of sampling. All four papers reporting surveys using *groups* for sampling *individuals* applied LinkedIn as source of sampling although clearly differing in their approaches. While JOORABCHI, MESBA and KRUCHTEN (2013) and KANIJ, MERKEL and GRUNDY (2011) sent indirect invitations using forum groups, our preliminary studies (DE MELLO and TRAVASSOS, 2013; DE MELLO et al., 2014) individually invited members from LinkedIn groups of interest retrieved by following specific recruitment plans, as mentioned in Section 3.3. All four surveys using *papers* for sampling *individuals* selected the papers used from the results of SLRs previously performed on each survey context (DIAS-NETO and TRAVASSOS, 2008; CARVER *et al.,* 2013; Santos and da Silva, 2013; GUZMÁN *et al.*, 2014), which has been showed an interesting alternative for establishing representative sampling frames composed by SE researchers.

Eight from the nine surveys using *organizations*/*organizational units* as search units were performed over sampling frames established by convenience, which reflects the challenge on accessing comprehensive sources composed by representative sets of SE organizations/ organizational units. Alternatively, CONRADI *et al.* (2005) and JI *et al.* (2011) try overcoming such challenge applying efforts on manually filtering IT Organizations working with COTS from nationwide generic sources of population (such as yellow pages) from four distinct countries (Italy, Germany, Norway and China).

Table 3-15. Surveys selected in the structured review [DE MELLO and TRAVASSOS, 2015).

Arrangement	#	Non-Convenience Sampling	Convenience Sampling
Individual/ Individual	19	(BEGEL and NAGAPPAN, 2008) (RODRÍGUEZ *et al.,* 2012) (MURPHY *et al.,* 2013)	(HOGGANVIK and STOLEN, 2005) (SLYNGSTAD *et al.,* 2006) (NUGROHO and CHAUDRON, 2007) (GUO and SEAMAN, 2008) (FRANÇA and DA SILVA, 2009) (NGUYEN-HOAN, FLINT and SANKARANARAYANA, 2010) (SOUTHEKAL and LEVIN, 2011) (TOFAN *et al.,* 2011) (TORCHIANO *et al.,* 2011) (KUSUMO *et al.,* 2012) (TOMASSETTI *et al.,* 2012) (CAVALCANTI *et al.,* 2013) (SENAPATHI and SRINIVASAN, 2014) (DIEBOLD *et al.,* 2014) (MONASOR *et al.,* 2014) (PFAHL, MÄNTYLÄ and MÜNCH, 2014)
Individual/ Group	4	(DE MELLO and TRAVASSOS, 2013) (DE MELLO, DA SILVA and TRAVASSOS, 2014) (2014b)	(KANIJ, MERKEL and GRUNDY, 2011) (JOORABCHI, MESBA and KRUCHTEN, 2013)
Individual/ Organization	4	-	(GLYNN, FITZGERALD and EXTON 2005) (ODZALY, GREER and SAGE, 2009) (FRANÇA, DA SILVA and MARIZ, 2010) (HUMAYUN, GANG and MASSOD, 2013)
Individual/ Paper	4	(DIAS-NETO and TRAVASSOS, 2008) (CARVER *et al.,* 2013) (SANTOS and DA SILVA, 2013) (GUZMÁN *et al.*, 2014)	-
Individual/ Project	1	-	(NUNNENMACHER *et al.*, 2011)
Organization/ Organization	4	-	(TAIPALE and SMOLANDER, 2006) ((TAIPALE, KARHU and SMOLANDER, 2007)) (KASURINEN, TAIPALE and SMOLANDER, 2010) (DIEBOLD and VETRÒ, 2014) (MOE *et al.*, 2014)
Project/ Organization	1	(CONRADI *et al.*, 2005) ((JI *et al.*, 2008))	-

Attributes applied on characterizing the surveys' units were also investigated, including those used to restrict the survey population (*control attribute*), such as the subject's *country* (CONRADI *et al.,* 2005; JI *et al.*, 2008; RODRÍGUEZ *et al.*, 2012). We observed the frequent use of the following attributes for characterizing *individuals*: *experience in the research context* (57%), *current professional role* (51%), *SE experience* (37%), *country* (35%), *professional status* (31%) and *academic degree* (29%). For characterizing *organizations*/*organizational units* we identified the following main attributes: *size* (78%), *industrial segment* (70%), *country* (65%) and *organization type* (48%). Among the few cases of units composed by observed *software projects*, we identified the use of several attributes, such as *project size and duration*, *software process applied*, *team size*, *client/ product segment*, *client nature* (public or private organization) and its *physical distribution*.

Although the findings obtained in the presented structured review evidenced the need on improving the quality of sampling activities in SE surveys, we could compile an initial set of recommendations for characterizing the unit of analysis and establishing the sources of sampling (DE MELLO and TRAVASSOS, 2015). Most of such recommendations was incorporated to the second version of the conceptual framework, presented in Chapter 4.

3.4.2. Investigating Recruitment and Participation

As already mentioned in the Introduction of this Thesis, efforts on improving sampling representativeness in SE surveys can be useless if the recruited individuals are not willing to participate. In general, we have observed *convenience* strategy in SE surveys is characterized by the following scenarios:

1. The adoption of a *non-probabilistic sampling design* (such as accidental sampling, quota, snowballing or judgment) (CAVALCANTI *et al.*, 2013);
2. The establishment of *very strict sampling frames*, restricting the survey sample to a non-representative population (although eventually large) from the point of view of the survey research objective (SLYNGSTAD *et al.*, 2006; FRANÇA and DA SILVA 2010; MOE *et al.*, 2014);
3. The *absence of a sampling frame*, observed when the survey recruitment is indirectly performed and there is no control on whose individuals can respond it (KANIJ, MERKEL and GRUNDY, 2011; JOORABCHI, MESBA and KRUCHTEN, 2013), and;
4. The *absence of a controlled sample*, observed when researchers allow the subjects to forward the survey invitation to other individuals, as observed the

convenience sample used in the survey re-executed by DE MELLO and TRAVASSOS (2013).

In all scenarios, it is possible to observe that subjects' participation can be influenced by some identification with the researchers/research group involved. In the scenarios "3" and "4", it will be not possible analyzing the convenience impact in the survey participation since we don't know how many subjects could participate and who they are. In general, we have observed participation rates in SE surveys using sampling by convenience (scenario 1 and/or scenario 2) higher than in those which sampling by convenience is avoided. Thus, we believe that such scenario has been influenced not only by the commitment of the subjects on participating but also by the scarce or even inadequate use of persuasion factors (SMITH et al., 2013) for stimulating subjects' participation. In order to better understand in which extent such factors and convenience have been influenced the voluntary participation in SE surveys, the following hypotheses emerged:

- *$H_0$1: There is no difference in participation rates between samples established by convenience and not established by convenience for surveys in SE.*
- *H_A1: There is difference in participation rates between samples established by convenience and not established by convenience for surveys in SE.*
- *$H_0$2: The use of persuasive factors in SE surveys does not influence survey participation rates.*
- *H_A2: There is at least one persuasive factor that influences participation rates in SE surveys when used.*

One can see that some other factors than those ones presented by SMITH et al. (2013) can influence participation in surveys, such as the *survey topic* and *target audience*. However, since the technical literature reports few survey re-executions, we decided to initially evaluate the presented hypothesis comparing surveys from different research topics and using different target audiences, using the same set of 37 surveys analyzed in Section 3.4. For supporting our analysis, the following data were collected from each survey: *sampling frame*, *sample size*, *effective sample size (number of participants)*, *invitation message* and *recruitment steps*.

Initially, we observed that relevant data for supporting our investigation were not available in most of technical papers. Thus, except by the three surveys conducted by ESE Group (DIAS-NETO and TRAVASSOS 2008; DE MELLO and TRAVASSOS 2013b; DE MELLO, DA SILVA and TRAVASSOS, 2014), all authors involved in the other 34 identified surveys were asked to provide further information regarding their surveys'

planning and execution. An e-mail was sent regarding each survey to each researcher, asking for a single reply. After 30 calendar days, we obtained additional information for 21 surveys.

Since the *participation rate* is the effect to be observed, 19 surveys reporting undefined samples (or without providing sufficient data about it) from the 37 in the sampling frame were excluded, remaining 10 surveys with *convenience samples* and 8 surveys having *non-convenience samples*. Thus, in order to balance the dataset, we added three other surveys having non-convenience samples conducted by the ESE Group (SANTA IZABEL and TRAVASSOS, 2011; SPÍNOLA and TRAVASSOS, 2012; ABRANTES and TRAVASSOS, 2013).

Table 3-17 presents all surveys' samples included in the analysis, ordered by participation rate. As MOE *et al.* (2014) applied different methods for recruiting two different groups in their survey, the samples were analyzed separately. On the other hand, DE MELLO, DA SILVA and TRAVASSOS. (2014a, 2014b) followed the same procedures to sampling and recruitment in all survey strata, therefore their samples were aggregated into a single one. **Error! Reference source not found.** shows the basic statistics of participation rates for convenience and non-convenience sampling. Both convenience and non-convenience distributions were observed as normal (Shapiro-Wilk test, log-normal distribution). Applying the Dixon's test for a single outlier in each distribution, the participation rate reported in (GUZMÁN *et al.*, 2014) was identified as outlier (p-value=0.007) and therefore removed from the non-convenience distribution. As expected, the standard deviation from the convenience distribution is higher. It can be explained due to variation of strategies used to introduce convenience on each survey.

Since Levene's test indicated that they were not homoscedastic distributions (p-value=0.008, confidence level=95%), Mann-Whitney non parametric test was applied to compare the distributions. We observed that participation rate in convenience samples was significantly higher than participation rate in non-convenience samples (p-value=0.0005, confidence level=95%). In addition, one can see through Table 3-17 that all 11 samples from the convenience distribution have target audiences composed by practitioners (varying on the research context involved and eventually including researchers), while five from the 10 samples from the non-convenience distribution have specifically *SE researchers* as target audience, which could hypothetically suggest higher participation rates. Coincidentally, all these five surveys followed the same strategy for sampling: recruiting authors of papers selected in SLR. However, it was not observed significant differences between participation rates of samples composed by researchers and practitioners in the surveys analyzed using non-convenience.

From the 21 surveys analyzed, seven did not report the application of a systematic recruitment process following a standard recruitment message. Thus, they were excluded from the evaluation of using persuasive factors. In general, we can observe that researchers' affiliation (*credibility*) is present in all surveys, while *compliment*, *direct invitation* and establishing a deadline for survey answering (*scarcity*) were factors commonly applied but the use of *reward* is seldom observed. BEGEL and NAGAPPAN (2008) and RODRÍGUEZ *et al.* (2012) enrolled respondents in raffles, while KUSUMO *et al.* (2012) present the only survey in the sample on which respondents were paid. In only two surveys (BEGEL and NAGAPPAN 2008; Moe et al., 2014) participation could be influenced by the similarity observed between one or more researchers involved and the respective survey sample. Finally, no incidence of using *humor* for stimulating subjects' participation was observed. The small sample analyzed and the diversity of arrangements of applied persuasive factors did not allowed us to perform statistical tests over the results. Thus, it was not possible to test $H_0 2$.

Table 3-16. Descriptive statistics of both analyzed distributions (number of outliers removed in parenthesis).

Distribution	Size	Mean	StdDev	Min.	Q1	Median	Q3	Max.
Convenience	11(0)	0.6609	0.2834	0.0900	0.5000	0.6538	0.9692	1.0000
Non-Conv.	11(1)	0.1318	0.0666	0.0368	0.0782	0.1252	0.1852	0.2361

Findings from the presented investigation suggest how the willingness of researchers/practitioners on contributing with SE research need to be improved in the field, apart from the research topic involved. It also strength our argument on how SE survey plans should be improved. One can argue that some great differences observed between participation rates from different surveys following non-convenience sampling could be influenced by the research topic involved. In fact, participation rates of surveys regarding diverse research topics (such as requirements, agile methods, programming, empirical software engineering) and target audiences (such as project managers, system analysts, programmers, SE researchers) were compared. Future investigations would include re-executions of a same survey on which different samples from the same population are recruited through following different sampling and recruitment strategies.

Table 3-17.The final set of surveys' samples analyzed, ordered by participation rate.

Sample	Convenience Sampling?	Subject Type	Sample Size	Effective Sample Size	Participation Rate
SLYNGSTAD *et al.*, 2006	Yes	Practitioners	16	16	100.00%
MOE *et al.*, 2014 (group 1)	Yes	Practitioners	18	18	100.00%
FRANÇA, DA SILVA and MARIZ (2010)	Yes	Practitioners	65	63	96.92%
DIEBOLD *et al.*, 2014	Yes	Practitioners	45	35	77.78%
FRANÇA and DA SILVA 2009	Yes	Practitioners	235	176	74.89%
MOE *et al.*, 2014 (group 2)	Yes	Practitioners	26	17	65.38%
KUSUMO *et al.,* 2012	Yes	Practitioners	111	69	62.16%
GUZMÁN *et al.*, 2014	No	Researchers	113	64	56.64%
NUNNENMACHER *et al.*, 2011	Yes	Practitioners	82	45	54.88%
GUO and SEAMAN, 2008	Yes	Practitioners	38	19	50.00%
DIEBOLD and VETRÒ, 2014	Yes	Practitioners	50	18	36.00%
DIAS-NETO and TRAVASSOS, 2008	No	Researchers	144	34	23.61%
ODZALY, GREER and SAGE, 2009	No	Practitioners	89	18	20.22%
ABRANTES and TRAVASSOS, 2013	No	Researchers	117	21	17.95%
BEGEL and NAGAPPAN, 2008	No	Practitioners	2,821	487	17.26%
CARVER *et al.*, 2013	No	Researchers	440	59	13.41%
SANTA IZABEL and TRAVASSOS, 2011	No	Researchers	172	20	11.63%
SPÍNOLA and TRAVASSOS, 2012	No	Researchers	280	31	11.07%
RODRÍGUEZ *et al.*, 2012	No	Practitioners	4,450	408	9.17%
CAVALCANTI e*t al.*, 2013	Yes	Practitioners	400	36	9.00%
DE MELLO, DA SILVA and TRAVASSOS, 2-2014 (2014b)	No	Practitioners	7,745	291	3.76%
DE MELLO and TRAVASSOS, 2013b	No	Practitioners	924	34	3.68%

3.4.3. Threats to Validity

We excluded some surveys from the presented analyses due to the absence of relevant data regarding sampling and/or recruitment activities. Distinguishing a survey as sampled by *convenience* or *non-convenience* could be considered a threat to validity and we tried to mitigate it by asking the researchers for more detailed information in the case of recruitment analysis. Although the notorious relevance of EASE and ESEM conferences to the field the presented investigation should be extended to other venues to provide more confident results. Specifically regarding the persuasive factors analysis, it is important to point out that more precise results should demand survey plans replicated several times using different arrangements of persuasive factors with different samples.

3.6. Conclusion

This chapter summarized the research steps conducted until the development of the second version of a conceptual framework to support the identification of representative samples in surveys in SE. Details of each research step can be found in the publications cited. Taking into account the lack of suitable sampling frames available to support sampling in SE, we started adapting concepts used in SLR to support the search and retrieving of representative samples available in generic sources. The positive results obtained using such concepts in the preliminary studies and the lessons learned from such studies allowed us to depict the first version of the conceptual framework (v1) addressed to sampling issues, despite not providing guidelines for supporting their use. In this sense, the findings from the structured review conducted drove us to evolve the conceptual framework to the second version presented in Chapter 4.

4 Conceptual Framework v2

In this chapter, we present the second version of the conceptual framework, detailing its second release and exemplifying its use.

4.1. Introduction

The experience gathered on applying the first version of the conceptual framework and the findings from the structured review presented in the Chapter 3 led us to evolve the conceptual framework to a second version (v2). Among other improvements, the statistical concept of *unit of analysis* was included while the concept of *recruitment strategy* was introduced to support the characterization of recruitment activities and their resources (DE MELLO and TRAVASSOS 2015b). The concept of *source of sampling* was renamed to *source of population* to better represent the immediate contribution of this concept: identify a representative and accessible set of the survey target audience, independent from the sampling activities. The concept of *measurement instrument* was replaced to *characterization questionnaire* because it is the measurement instrument commonly observed in survey research (DE MELLO AND TRAVASSOS, 2015). Table 4-1 shows the conceptual framework concepts and its respective properties, highlighting the differences from the first version. The complete definition of each concept (with a complete example) is presented in Section 4.2.

The conceptual framework v2 also introduced a set of survey planning *activities* composed by *tasks* and *recommendations* designed to guide researchers to applying its concepts. The BPMN model presented in Figure 4-1 highlights the six survey planning *activities* (white boxes) designed and how they are distributed into the whole survey planning process, based on KASUNIC's (2005) survey steps. Since the *research objectives* are identified, a *target audience* should be established and a suitable *sampling frame* should be characterized by executing a *population search plan* over the selected *source of population*. Then, a compatible *sampling strategy* should be applied in order to deliver the survey sample. Finally, after the survey *questionnaire composition*, the *recruitment strategy* should be designed. For each activity, one or more *tasks* were developed (17 in total). Each task may be supported by one or more *recommendations* (27 in total).

Table 4-1. Comparison between conceptual frameworks v1 and v2.

Conceptual Framework v1		Conceptual Framework v2	
Concepts	**Properties**	**Concepts**	**Properties**
Target Audience	-	Target Audience	-
Unit of Observation	Unit of Observation attributes	Unit of Observation	Unit of Observation attributes
-	-	Unit of Analysis	Unit of Analysis attributes
Source of Sampling	-	Source of Population	-
Search Unit	Search Unit Attributes	Search Unit	Search Unit Attributes
Search Plan	Search String	Population Search Plan	Search String
	Search Algorithm		Search Algorithm
	Exclusion Criteria		Exclusion Criteria
Sampling Frame	-	Sampling Strategy	Sampling Frame
Sampling Strategy	Sampling Design, CL, CI		Sampling Design
-		Recruitment Strategy	Execution Estimated Time
			Invitation Method
			Period Available
			Reminding Method
			Reward Method
Measurement Instrument	-	Characterization Questionnaire	-

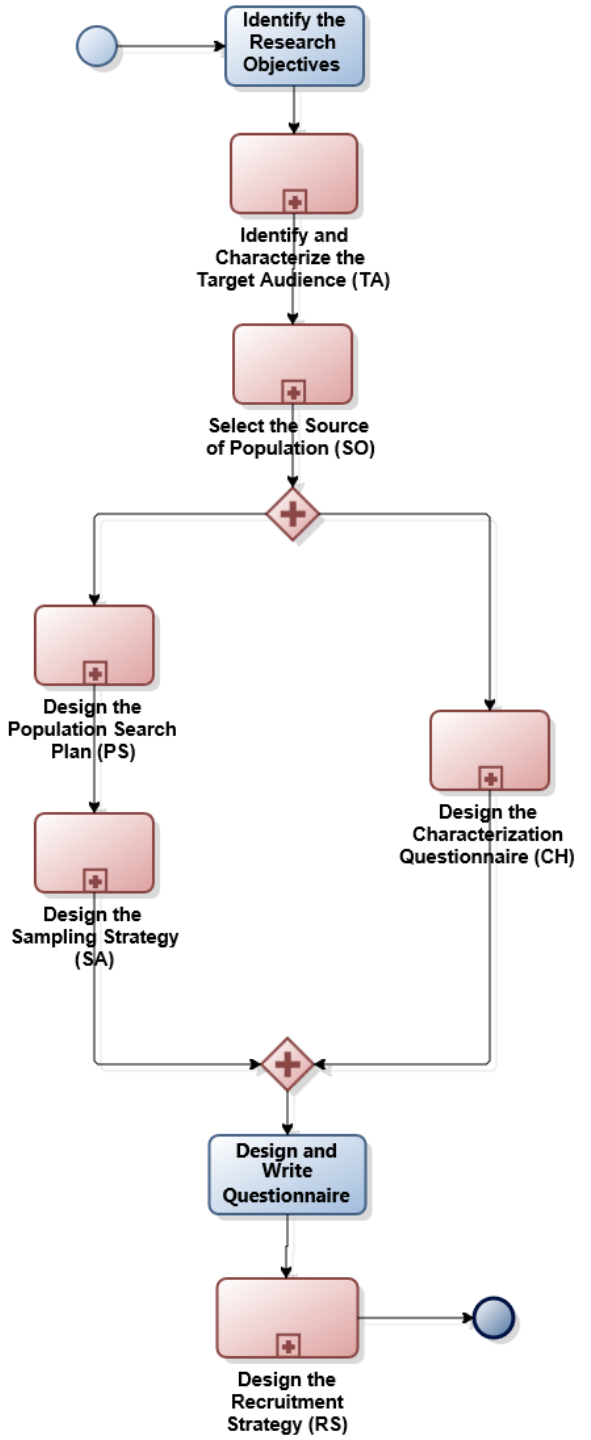


Figure 4-1. The framework activities inserted in the survey planning process.

A first release of the conceptual framework v2 (v2.1) containing a complete example of a survey plan instantiating the framework concepts was submitted to a *proof of concept* (6) in which a doctoral student at COPPE/UFRJ (external to ESE group) was invited to read the conceptual framework documentation and apply it to plan a survey on

features on software development processes, addressed to her Thesis. It was the subject first experience on planning a survey but she had already designed the survey questionnaire before the proof of concept. After applying the conceptual framework v2.1 the subject was invited to answer a follow up questionnaire regarding the technology acceptance. As a result, we observed the conceptual framework was considered useful and easy to use although we observed few concepts being wrongly instantiated, which could be explained by issues found in the description of some framework concepts and tasks. In addition, we could observe that the survey plan example presented in the conceptual framework probably had influenced her choices. Then, we made improvements in such items, generating a new release (v2.2) preserving the concepts and activities presented in Table 4-1 and Figure 4-1 but removing the survey plan example. Sections 4.2 and 4.3 present the conceptual framework v2.2 concepts and activities, respectively. Section 4.4 presents an example of using the conceptual framework.

4.2. Conceptual Framework v2.2: Concepts

4.2.1. Target audience

A survey *target audience* characterizes who can best provide the information needed in order to achieve the *research objective* (KASUNIC et al., 2005).

4.2.2. Unit of observation and unit of analysis

In opinion surveys, data is always collected from *units of observation* (primary objects) necessarily represented by the individual (respondent) (Linåker, 2015). However, the survey target audience may demand a level of analysis (*unit of analysis*) distinct from a single individual, including *organizational units, organizations and project teams* (DE MELLO and TRAVASSOS, 2015). For instance, CONRADI *et al.* (2005) conducted a survey in which individuals (unit of observation) that worked in software projects (unit of analysis) applying off-the-shelf (OTS) components was the target audience.

4.2.3. Units' attributes

Survey unit of observation and unit of analysis should be characterized through a set of *attributes*. *Control attributes* followed by predefined values can be used for representing the restrictions previously established by the target audience. For instance, if the target audience restricts the survey to be applied to *Brazilian individuals working*

as project managers then *"country= Brazil"* and *"role= project manager"* are control attributes. In addition, another attributes may be used to support the survey results analysis. DE MELLO and TRAVASSOS (2015) identified that individuals are commonly characterized through their *experience on the research context* and their current *organizational role*, while organizations are commonly characterized through their *size*, *industry segment* and *country*.

4.2.4. Source of Population

Survey population consists on a set of accessible units of analysis from a specific target audience (THOMPSON, 2012). Thus, a *source of population* consists on a database (automated or not) from which an adequate population for specific target audience can be systematically retrieved. If a source of population can be considered *valid* in a specific research context, it can be concluded that *sampling frames* can be established from it for the same research context.

4.2.4.1. Search Unit

The *search unit* characterizes the entity from which one or more units of analysis can be retrieved from a specific *source of population*. Indeed, in an ideal scenario, it is expected that both *unit of analysis* and *search unit* are the same thing (ESEM14-FW). However, SE literature presents some examples in which these units are different, such as the following:

- CONRADI *et al.* (2005) aims at analyzing project teams, but accessed them keeping in touch with *organizations* from three distinct countries.
- DIAS-NETO and TRAVASSOS (2008), CARVER *et al.* (2013), SANTOS *et al.* (2013), ABRANTES and TRAVASSOS (2013) and GUZMÁN *et al.* (2014) opted to survey the authors of the *papers* retrieved from the results of specifics SLRs conducted for each research context.
- DE MELLO, DA SILVA and TRAVASSOS (2015) used *groups of interest* from the professional social network LinkedIn to sample individuals, since the tool significantly restricted the direct access to individuals out from their groups.

Figure 4-2 associates the concepts of *source of population* and *search unit* (SU) with the concepts of *target audience*, *population* and *unit of analysis*. One can see that not necessarily all instances of search unit from a source of population can be used to compose a specific survey population.

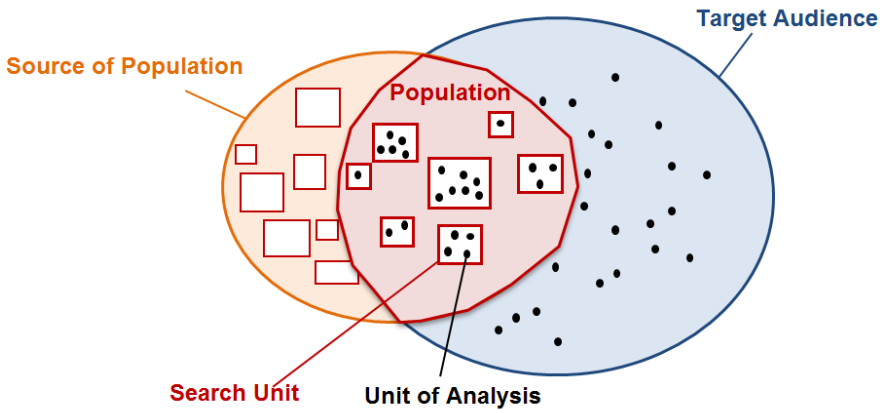


Figure 4-2. Population identified from a hypothetical source of population, considering a hypothetical target audience.

4.2.4.2. Criteria for Assessing Sources of Population

To be considered *valid*, a source of population should satisfy, at least, the following *essential requirements* (ER) (DE MELLO et al., 2014):

- ER1. A source of population should not intentionally represent a segregated subset from the target audience, i.e., for a target population audience "X", it is not adequate to search for units from a source intentionally designed to compose a specific subset of "X".
- ER2. A source of population should not present any bias on including on its database preferentially only subsets from the target audience. Unequal criteria for including search units mean unequal sampling opportunities.
- ER3. All source of population' search units, their units of analysis (and their units of observation) must be identified by a logical or numerical id.
- ER4. All source of population' search units must be accessible. If there are hidden search units, it is not possible to contextualize the population.

There are also nine desirable requirements (DR), three concerned with the samples' accuracy (ADR), two concerned with clearness (CDR) and four regarding sample's completeness (CoDR).

- ADR1. It is possible to retrieve each search unit from the source of population in a logical and systematic way.
- ADR2. There are no units of analysis outside the target audience concerned with the source of population.
- ADR3. There is a one-to-one correspondence between each search unit and each unit of analysis of the target population.

- CDR1. All search units appear once in the source of population.
- CDR2. All units of analysis appear once in the source of population.
- CoDR1. All information needed from each search unit is up-to-date.
- CoDR2.All information needed from each unit of analysis and their units of observation is accessible and up-to-date.
- CoDR3.All units of analysis from the target audience can be found in the source of population.
- CoDR4. Each search unit provides relevant information for performing alternative probabilistic sampling designs, such as stratified and cluster sampling. The source of population directly supports the researcher on classifying and/or clustering its population.

4.2.4.3. Sources types

Depending on the survey unit of analysis and the research context, many types of sources can be used. DE MELLO and TRAVASSOS (2015) identified in the SE literature the following types of source of population (search unit in italic):

- *SE Conferences: individuals* assisting to relevant SE conferences can be eventually considered a small but representative set for many surveys contexts, especially when researchers are the target audience (TORCHIANO *et al.*, 2011; MONASOR *et al.*, 2014);
- *Discussion Groups:* an active and thematic SE discussion group can be considered as good source for sampling *individuals* (NUGROHO and CHAUDRON, 2007);
- *Projects repositories:* retrieving a representative dataset from *software projects*, including data from the project team, is a challenge. Typically, data from several projects can be retrieved in the context of open source projects (BETTENBURG *et al.*, 2008).
- *Digital Libraries:* when it is expected that units of analysis are restricted to researchers, Digital libraries such as SCOPUS and IEEE can be used for retrieving relevant authors of *papers* in the survey context (DIAS-NETO and TRAVASSOS., 2008; CARVER *et al.*, 2013; SANTOS and DA SILVA, 2013; ABRANTES and TRAVASSOS, 2013; GUZMÁN *et al.*, 2014);
- *Catalogues:* searching for National or International catalogues provided by institutes (RODRÍGUEZ *et al.*, 2012), government or even yellow pages (CONRADI *et al.*, 2005) may be considered for retrieving representative sets of *organizations* or *individuals*;

- *Professional Social Networks:* it has demonstrated to be a promising technology for supporting large scale sampling of *individuals*, that can be directly accessed or through *groups of interest* (DE MELLO, DA SILVA and TRAVASSOS, 2015; TORCHIANO *et al.*, 2011). However, limitations on accessing searching units must be taken into account;

This list does not intend to be exhaustive and other types of sources can be taken into account. DE MELLO et al. (2014) applied the essential requirements and the desirable requirements in nine distinct sources of sampling available in the Web, including *crowdsourcing tools*, *professional social networks* and *freelancing tools*. While it was observed that the analyzed freelancing tools and professional social networks could be used as source of population, it was concluded that the analyzed crowdsourcing tools do not support to essential requirements.

4.2.5. Characterization Questionnaire

Attributes needed for characterizing each individual are frequently unavailable in the sources of population. Thus, such attributes data are commonly retrieved based on subjects' answers to one or more survey questions. For instance, DIAS-NETO and TRAVASSOS (2008) directly asked to the individuals all the attributes needed, since the source of population/search unit used (digital libraries/papers) allowed to access only individuals' names and their e-mails. In the same way, ABRANTES and TRAVASSOS (2013) asked to each individual about their experience with agile development, since it was not possible to collect such information from all members at LinkedIn profiles. The characterization questionnaire are typically included at the beginning or at the end of the survey questionnaire and it should be avoided to ask any information regarding the individual that is already available (and updated) in the source.

4.2.6. Population Search Plan

A *population search plan* describes how instances from the survey search unit will be systematically retrieved from the selected source of population and evaluated in order to compose the survey population (DE MELLO et al., 2014). The following subsections present the population search plan properties.

4.2.6.1. Search String

A *search string* is composed by a set of *search expressions* connected through logical operators that can be applied to a source of population in order to retrieve

adequate search units. As in the case of systematic literature reviews (SLRs), we argue that search expressions can be applied to avoid bias on filtering the relevant elements from the point of view of the *research objective*. Search strings must be avoided when there are no units of analysis outside the target audience concerned with the source of population (ADR2). This could happen, e.g., when the source of population is composed of the list of employees from a SE organization, and the set of employees from this organization composes the target audience.

4.2.6.2. Search Algorithm

The *search algorithm* describes each step, automated or not, that should be followed in order to filter the search units in a source of population, including how to apply the planned search string. A search algorithm can vary significantly on complexity, depending on the resources available in the source of population. In addition, any previously known restrictions for accessing the search units (and how to deal with them) should be described.

4.2.6.3. Exclusion Criteria

Another concept borrowed from SLRs, the *exclusion criteria* describe a set of restrictions that should be applied in order to exclude undesirable search units retrieved from the search plan execution. Exclusion criteria can be especially helpful when the source of population is significantly generic and the use of search string is limited, such as in the case of the professional social networks (DE MELLO, DA SILVA and TRAVASSOS, 2015) and yellow pages (CONRADI *et al.*, 2005). As in the case of search strings, if the requirement ADR2 is satisfied by a source of population, exclusion criteria should be avoided.

4.2.7. Sampling Strategy

A *sampling strategy* establishes criteria for composing the survey *sampling frame* and describes the survey *sampling design*.

4.2.7.1. Sampling frame

In statistics, a *sampling frame* is the source from which a sample, i.e. a subset of units from the population can be retrieved (SÄRNDAL, SWENSSON and WRETMAN, 1992). 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. SÄRNDAL, SWENSSON and WRETMAN (1992)

observed that some appropriate investigations could not be carried out due to the lack of a suitable sampling frame, while other investigations remain with inconclusive results due to incomplete sampling frames. Frequently, all accessible population from each SE survey is used to compose the survey sampling frame.

4.2.7.2. Sampling design

Describes the criteria for extracting samples from the sampling frame, i.e., which individuals (from each unit of analysis) will be invited to answer the survey. In general, probabilistic (randomly) sampling approaches are recommended (LINÅKER *et al.*, 2015): *simple random sampling*, *clustered sampling*, *stratified sampling* and *systematic sampling*. Eventually, all individuals in the sampling frame can be included in the sample (census).

4.2.8. Recruitment Strategy

The recruitment strategy characterizes how the individuals from the survey sample will be recruited. It includes the *invitation message* and the following factors that can influence subjects' participation (SMITH *et al.*, 2013).

- *Execution estimated time:* the mean time estimated to each subject fill out the survey. It can be calculated based in the results of pilot executions (KASUNIC, 2005);
- *Invitation method:* characterizes how the invitation message will be sent. In the case of invitations supported through Web, common approaches are *sending individual and personalized e-mails; individual and generic e-mails; sending a single generic e-mail; sending a single generic message to a list or a group*.
- *Period Available:* characterizes how for many times (typically in days) the survey will be available for the subjects.
- *Reminding method:* any intention of reminding individuals regarding answering the survey should be described. For instance, it can be planned to remind the subjects once, re-sending the original message after one week of the survey execution.
- *Reward method:* it must be described if the subjects can be stimulated through any kind of reward and in which case it will be offered (e.g. if only the subject complete the survey questionnaire). In this context, rewards may but is not limited to include *payments*, *raffles*, *gifts* and *donations for NGOs*.

4.3. Conceptual Framework- Activities

4.3.1. Identify and Characterize Target Audience (TA)

Since the survey research objectives are already established, the researchers should be able to identify and characterize its target audience.

TA01. *Identify the survey target audience.* Based on the research objectives, answer the following question: *"Who can best provide you with the information you need?"* (KASUNIC, 2005). Thus, avoid restricting the target audience based on factors such as its size or availability.

TA02. *Identify the unit of analysis.* Identify from which arrangement of individuals it is expected to analyze and interpret the survey results.

R1. It has been observed that the own *individual* is more commonly used as unit of analysis in SE surveys, followed by *organizations*, *organizational units* and *project teams*.

TA03. *Establish the unit of analysis attributes.* First, establish the set of control attributes and their respective values that will restrict the unit of analysis. Then, enumerate the other attributes that should be collected from each unit and define how to measure each one.

R2. Individuals are commonly characterized in SE through the following attributes (DE MELLO and TRAVASSOS, 2015): *experience in the research context, experience in SE, current professional role, country* and *higher academic degree*.

R3. Organizations are commonly characterized in SE through the following attributes (DE MELLO and TRAVASSOS, 2015): *size (scale typically based in the number of employees), industry segment (software factory, avionics, finance, health, telecommunications, etc.), country and organization type (government, private company, university, etc.*

R4. Project teams can be characterized through attributes such as *project size; team size, client/product professional segment (avionics, finance, health, telecommunications, etc.)* and *physical distribution* (DE MELLO and TRAVASSOS, 2015).

R5. Taking into account the limitations of each data type involved, establish formulas to measure characteristics composed by two or more attributes. For instance, DE MELLO, DA SILVA and TRAVASSOS (2015) calculated the subject's experience level through a weighted mean between four distinct dimensions: highest academic degree (scale); number of years working with SE; number of years working with the research scope; number of SE projects already concluded (scale).

R6. When possible, follow already established standards to support the measurement of attributes. Standards can be especially helpful in the case of attributes measured through scales and nominal variables.

TA04. *Establish the unit of observation attributes.* This step should be performed whether the unit of analysis is not represented by the same entity as the unit of observation, i.e., the individual. Again, recommendations R02, R05 and R06 are applicable.

4.3.2. Select the Source of Population (SO)

This activity aims at establishing the source from which is expected to identify an accessible population from the survey target audience.

SO01. *Identify the Source of population candidates.* If not all target audience are accessible through a single and specific source (a common limitation in SE research), you must look for alternative sources and their respective search units available that could retrieve a representative subset from them.

R7. Avoid the convenience on searching sources. Have in mind to find sources that could provide scalability and heterogeneity on the research context.

R8. If the target audience is limited to SE researchers (or subsets from them), valuable sources of sampling includes digital libraries such as SCOPUS and IEEE while the search unit can be the papers available in such sources. Professional social networks such as ResearchGate and Academia.edu can be also useful for searching SE researchers.

R9. Look for catalogues provided by recognized institutes/ associations/ local government to retrieve relevant set of SE professionals/ organizations. For instance, SEI institute provides an open list of the organizations and

organizational units certified in each CMMI-DEV level. FIPA provides information regarding Finland IT organizations and its professionals. CAPES (from the Brazilian government) provides a tool for accessing information regarding Brazilian research groups.

R10. The professional social network LinkedIn have been showed a useful source to access representative samples composed by SE professionals through their groups of interest.

R11. Other types of possible Source of population includes, but are not limited to: open discussion groups, SE conferences, project repositories and freelancing tools.

SO02. *Select the Source of population.* Discard any combination of Source of population candidate and its search unit that do not support the essential requirements (ER). Then, apply the desirable requirements (DR) to support your decision on selecting the survey Source of population.

R12. Before take your decision, invest efforts on simulating the use of each source and its search unit, since some technical limitations may be not explicit in a first glance. If you are not sure in which Source of population should be selected, consider run a pilot trial using each one to support your decision.

R13. Report any special condition needed to use the selected Source of population. For instance, a LinkedIn "Premium" account was needed to make feasible some group analysis identify members in common (overlapping) between different groups of interest.

SO03. *Establish the search unit attributes.* If the search unit is not represented by the same entity of the unit of analysis, identify relevant attributes to characterize each instance of the search unit. They can be helpful on filtering the survey population from a generic Source of population.

4.3.3. Design the Population Search Plan (PS)

This activity proposes a set of steps to extract the survey population from the source of population.

PS01. *Design the search string.* The use of search strings can be helpful on filtering the search units relevant to the survey context. Depending on the coverage of the Source of population and on its filtering resources, even complex searches may be needed, using logical operators as commonly performed in Systematic Literature Review (SLR).

R14. Consider consulting the specialized literature (especially standards such as IEEE vocabulary for Software Engineering) for identifying a wide range of relevant and similar expressions for composing the survey search string. If a SLR was previously performed in the context of your research, consider reusing its own search strings and the strings provided by its results.

PS02. *Design the search algorithm.* A search algorithm must be designed to describe the steps on applying the search string.

PS03. *Design the exclusion criteria.* Design the exclusion criteria that will be used to qualitatively filter the results from the execution of the search algorithm.

R15. Avoid introducing selection bias in the exclusion criteria and establishing subjective and ambiguous exclusion criteria. Evaluate if the exclusion criteria includes conditions that can be automatically verified. If so, in order to avoid introducing operational errors, remove it and update the search algorithm/ search string.

4.3.4. Design the Sampling Strategy (SA)

This activity aims at describing how it is expected to extract a sample from the survey population.

SA01. *Establish the sampling frame.* The sampling frame can be composed by all the search units available in the survey population, resulted from the execution of the previous activities. However, there are situations in which the use of the whole population as sampling frame may imply in a disproportional or even unnecessary research effort. In such cases, rules to restrict the sampling frame to a subset of the population may be applied.

SA02. *Establish the sampling design.* A single sampling design should be selected. Statistics literature presents the more common probabilistic sampling designs and formulas for calculating survey sample size (THOMPSON, 2012).

R16. Avoid introducing any non-probabilistic step in the sampling design. For instance, in case of subjects being selected from a sample composed by a set of SE organizations, it is important to avoid the introduction the bias of organization representatives on select them.

R17. If there is enough information in the units attributes that clearly allows you to extract homogenous subpopulations for supporting clustered sampling, evaluate the feasibility of applying stratified sampling.

R18. When calculating the sample(s) size(s), have in mind that participation rates in voluntary SE surveys over random samples tend to be small (less than 10% of the sample size).

4.3.5. Design the Characterization Questionnaire (CH)

CH01. *Identify the unavailable individuals' attributes.* Identify which individual' attributes needed are not accessible or even not updated in the Source of population.

CH02. *Compose the Characterization questions.*

R19. Simple and optional open questions can bring an additional contribution to trace subjects' profile without overloading them with several questions/options, especially when there are few data available and updated in the Source of population. However, the coding process in large scale can be exhaustive and it must be considered in the study planning (DE MELLO, DA SILVA and TRAVASSOS, 2014b).

4.3.6. Design the Recruitment Strategy (RS)

This activity aims at establishing how to recruit the sample for participating in the survey.

RS01. *Establish the invitation method.* Considering the Source of population and the individuals' contact data available, establish how each sample subject will be invited to answer the survey.

R20. Samples in SE surveys are often invited through e-mails. In such case, take preference to send individual and personalized messages using a standard message as a template. If possible, activate e-mail resources to notify you whether the message was read by each subject.

R21. In the case of the survey samples are distributed through discussion groups/ groups of interest in the Web, avoid posting a generic forum invitation message, except if you have control about who will see it and how many members will be notified about it.

R22. Depending on the source of population used, no contact data may be available. Alternatively, professional social networks may allow researchers to keep in touch with the subjects using internal message services. However, have in mind that such sources may restrict the use of these services in large scale.

RS02. *Parameterize the recruitment.* Establish values to the recruitment parameters presented in the concept “Recruitment Strategy”:

R23. A finite and not so long period to answer the survey allows subjects to plan themselves about when they will answer the survey (scarcity).

R24. It is common in SE surveys to send a single remind message. Sending too many reminds in a short period, especially when the subjects are not previously compromised to collaborate (common in SE research), can bring a negative effect in the participation rate and in the quality of the survey data.

R25. It is expected that rewards can stimulate the participation in surveys, but such practice is still uncommon in SE research. Rewards can be offered only to the subjects who successfully completed the survey questionnaire or even to all subjects. In international surveys, it can be hard to establish a low-cost payment instrument for small rewards. Alternatively, donations can be offered or even a raffle can be performed between the participants.

RS03. *Compose the invitation message.* The invitation message should clearly characterize the researchers involved, the research context and present the recruitment parameters.

R26. In individual messages, introduce persuasive factors that can stimulate the subject participation, such as a compliment and an observation regarding the relevance of each subject participation.

R27. In on line surveys, it is highly recommended to send an exclusive token for each participant in order to avoid noise on repeated participations or even the unauthorized distribution of the survey.

4.4. Example

Table 4-2 presents the instantiation of the conceptual framework to support a survey on SLR usage in SE research, not yet conducted. The survey research objective is to characterize in which extent postgraduate students have been applied SLR to support their investigations and what are their perceived expectations/impressions regarding such method. The set of Brazilian SE conferences' proceedings and the list of research partners from the ESE Group were also candidates to source of population. However, CNPq (Brazilian council for scientific and technological development) research group directory was observed as the best option since it supports all essential requirements and the most of desirable requirements (Table 4-3).

4.5. Conclusion

This chapter presented the conceptual framework v2.2, evolved from a previous release (v2.1) submitted to a proof of concept. Different from the first version of the conceptual framework, the second version guides the use of its concepts by introducing a set of activities, tasks and recommendations for applying its concepts in the context of the survey planning process. The resultant survey planning process takes into account a (common) scenario in the field, in which there is no suitable sampling frame immediately available addressed to the survey target audience.

One can see that many recommendations provided by the conceptual framework v2.2 are grounded on the lessons learned on conducting the previous studies presented in sections 3.3 and 3.4, while others are grounded on the small set of interesting reports observed in the technical literature (Section 3.5). In fact, we expect that the evolution of survey research in the field can provide a more comprehensive and evidence-based set of recommendations in the future. Chapter 5 presents the empirical evaluation of the conceptual framework v2.2 through a feasibility study and a focus group session.

Table 4-2. Example of instantiating the conceptual framework v2.2 to support a survey on SLR usage in SE.

Concept	Property	Description
Target Audience	-	Brazilian SE Research Groups
Unit of analysis	Type of Entity	Research Group
	Attributes	Control attributes: *location= "Brazil" and research field= "SE"* Other attributes: *name, description, location (city), age, number of doctors, number of doctoral students and number of master students*
Unit of Observation	Attributes	Higher Academic Degree= *"PhD", "DSc" OR "Master"* Other attributes: *researcher name, time in the research group (years), experience with SE research, experience with SLR*
Source of Population	Description	CNPq research groups directory (http://dgp.cnpq.br/dgp/)
	Search Unit	Research Groups
Population Search Plan	Search String	*"software engineering"* OR *"engenharia de software"*[11]
	Search Algorithm	Apply each search string (between quotes) once in the field *"Search term"*, choosing the options *"Exact search"* and *"search by group".* Check to apply each search only to the following fields: *"group name", "research line name"* and *"keyword of the research line"*. Do not apply other filters available.[12]
	Exclusion Criteria	To exclude all research group retrieved created only after 2010 To exclude any research group in which description do not clear mention the use of empirical (experimental) methods to support their research in SE
Sampling Strategy	Sampling Frame	The list of all research groups resulted from the execution of the Population Search Plan
	Sampling Design	Simple Random Sampling of the research groups available in the sampling frame, selecting all the subjects available in each group.
Recruitment Strategy	Description	Since the e-mail of the subjects is not available in the source, to send a standard e-mail to the leader from each research group, asking him/her redirect the survey invitation

[11] "Software Engineering" in Portuguese

[12] The actual names of the fields are in Portuguese

	Invitation Message	*To be described*
	Estimated Execution Time	*Not estimated*
	Period Available	10th January 2016 - 24th January 2016 (15 days)
	Reminding Method	In the 7th day of survey execution, to re-send the invitation message only to the leaders of those research groups that no response was received
	Reward Method	None
Characterization Questionnaire	-	*It should include the following single question: "How long (years) you have been working with SE research?"*

Table 4-3. Evaluation of the candidates to source of population, based on the conceptual framework requirements.

Candidate	ER1	ER2	ER3	ER4	ADR1	ADR2	ADR3	CDR1	CDR2	CoDR1	CoDR2	CoDR3	CoDR4
Research Partners	×	✓	✓	✓	-	-	-	-	-	-	-	-	-
Conferences Proceedings	✓	✓	✓	✓	✓	×	×	✓	×	×	×	×	×
CNPq Research Groups Directory	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓

5 Empirical Evaluation of the Conceptual Framework v2.2

This chapter presents the empirical evaluations conducted for investigating the contributions of the conceptual framework v2.2 on supporting survey planning activities.

5.1. Introduction

Based on the positive results observed in the last evaluations of the conceptual framework, we decided to perform the empirical evaluation of the conceptual framework v2.2 by conducting the feasibility study described in Section 5.2. As a result, it was evidenced few effective contributions of the proposed technology to the quality of the survey plans elaborated, although the conceptual framework users and readers tend to accept it. Therefore, Section 5.3 presents a focus group section conducted to investigate in depth opportunities for improving the conceptual framework based on the feasibility study results. The content of this chapter is heavily based on (DE MELLO and TRAVASSOS, 2016).

5.2. Feasibility Study

Since we are introducing a new research technology into the context of survey planning, we are concerned with investigating its potential contributions to the quality of the studies as well as to investigating the technology acceptance by its potential users. Thus, we established the following research goals:

- RG1: *To analyze* the development of survey plans supported/ not supported by the conceptual framework *in order to* characterize *with respect to* their thoroughness *from the perspective of* SE researchers *in the context of* evaluating survey planning items designed by other SE researchers.
- RG2: *To analyze* the experience on reading and/ applying the conceptual framework *in order to* characterize *with respect to* its perceived acceptance and the perceived relevance of the conceptual framework recommendations *from the perspective of* SE researchers *in the context of* answering a follow up

questionnaire regarding the experience on using/reading the conceptual framework.

Regarding *RG1*, the following null/alternative hypotheses emerged:

- H_A1 ($H_0$1): *There is (no) difference between the thoroughness of survey plans designed by using and not using the conceptual framework.*
- H_A2 ($H_0$2): *There is (no) difference between the thoroughness of survey plans characteristics designed by using and not using the conceptual framework.*

To support the testing of H_01 and H_02 we conducted a controlled experiment in which students from an Empirical Software Engineering course at COPPE/UFRJ were invited to individually fill in a survey plan on factors influencing effort on requirements engineering. Since the conceptual framework is not designed to support the whole survey planning process, the experiment instrument was composed by an actual but incomplete survey plan, having only the original survey *research objective* and *questionnaire* already defined. In order to support comparison, a subset from the subjects should perform the experimental task applying the conceptual framework v2.2 (FWK) while the other one does not (ADH). Each subject should be requested to report the following survey plan items: *target audience*, *participants' attributes*, *sampling frame*, *sampling design* and *recruitment process*. The survey plan characteristics to be evaluated were adapted from the *thoroughness items* and t*rustworthiness attributes* proposed by STAVRU (2014)[13] for evaluating survey reports. Regarding *RG2*, the following hypotheses emerged:

- $H_0$3: *There is no perception on the acceptance/non-acceptance of the conceptual framework.*
- H_A3.1: *There is perception on the usefulness of the conceptual framework.*
- H_A3.2: *There is perception on the easiness of use of the conceptual framework.*
- H_A3.3: *There is perception on the intention on using the framework.*
- $H_0$4: *There is no difference on the perceived relevance of the recommendations of the conceptual framework.*
- H_A4.1: *There are recommendations of the conceptual framework perceived as more relevant than others.*

[13]We added the thoroughness item *participants' attributes* to the original set of attributes and adapted the *target population (audience)* item, including the *unit of analysis*.

To support testing the hypotheses addressed to RG2, the subjects concluding the survey planning tasks will be requested to answer a *follow-up questionnaire*, including questions regarding the perceived conceptual framework acceptance (usefulness, easiness of use, intention to use) and a question addressed to the perceived relevance of its recommendations. Once ADH subjects will not have used the conceptual framework in the experimental task, they will be oriented to first fully read it to then answering the follow up questionnaire.

Table 5-1. Thoroughness items and trustworthiness attributes applied in the survey plans' evaluation. Adapted from (STAVRU, 2014).

Thoroughness item	**Description**	**Trustworthiness attribute**
Target Audience/ Unit of Analysis	The study specifies and thoroughly describes its target audience and unit of analysis	Truth value
		Consistency
Participants attributes	The study specifies the attributes that will be used to restrict the composition of the survey sampling frame and to characterize each participant	Truth value
		Consistency
Sampling frame	The study specifies and thoroughly describes its sampling frame, including from which source it will be obtained	Applicability
		Consistency
Sampling design	The study specifies and thoroughly describes its sampling design	Applicability
		Consistency
Recruitment Process	The study describes how the survey will be mediated	Consistency
	The study specifies how long the survey will be available to respondents	Consistency
	The study specifies how the invitations will encourage response and prevent non–response	Consistency

5.2.1. Experimental Design

The Subjects will be evenly distributed between the two groups balanced by their experience in the survey topic. All subjects will be submitted to a 4-hour survey class. In this class, all steps of the survey process shall be presented and discussed without mentioning the conceptual framework or its specific components. After the class, each subject will be individually invited by e-mail to perform the experimental task (Appendix A), receiving the survey plan form and the guidelines for conducting surveys in SE (LINÅKER *et al.*, 2015). In addition, FWK subjects should be oriented to follow the conceptual framework v2.2, receiving its documentation. Although the experiment tasks will be performed in Portuguese, the part of the experimental task document describing the survey research objective and the survey questionnaire was copied from the original survey plan, in English (VAZ, 2013). After a week, subjects that completed their

experimental task will be invited to answer a follow-up questionnaire. A specific questionnaire was designed to each group (although presenting some questions in common): FWK questionnaire was driven to ask subjects experience on applying the conceptual framework, while ADH questionnaire was driven to ask subjects on how the conceptual framework could affect the survey planning activity. Thus, ADH subjects will receive the conceptual framework documentation in the follow-up invitation.

5.2.2. Execution

The study was conducted in July 2015. All 12 students were invited to participate: 4 doctoral students, 6 master students and 2 candidates to be doctoral students. All of them declared no previous experience on conducting empirical studies in SE and different experience levels on the survey theme (software requirements effort influence factors). Since 11 subjects answered the consent form, six of them were randomly assigned to compose the FWK group while the five other subjects composed ADH, as presented in Figure 5-1. The numbers in parenthesis along each subject identifier represents its calculated experience rate in the survey theme (maximum is 1).

Nine subjects performed the experimental task, delivering fulfilled survey plans: five from ADH and four from FWK. Then, all of them answered to the follow up questionnaire. Following subsections presents the results and the hypotheses testing by research activity.

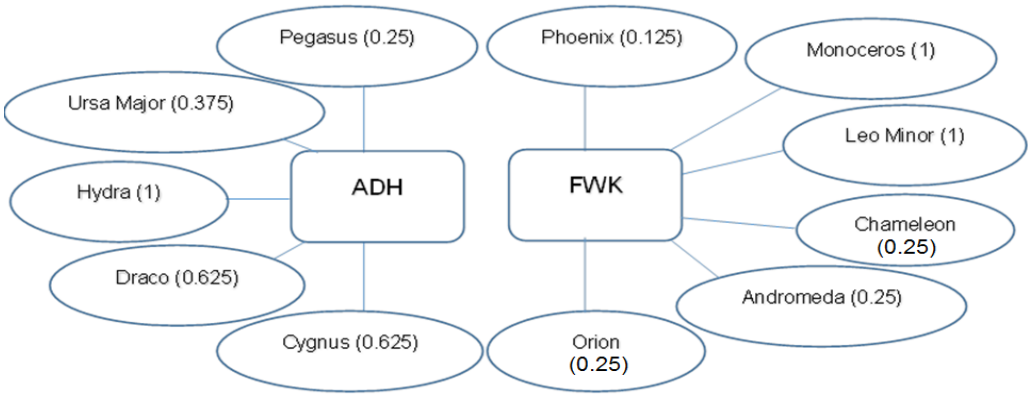


Figure 5-1. Composition of ADH and FWK groups

5.2.3. Experimental Task Results and Analysis

Two other experienced researchers from the ESE Group individually evaluated the thoroughness of the survey plans. The plans were forwarded to them without the identification of subjects. For each survey plan characteristic, researchers individually indicated whether the subject's response to a survey plan item brings a positive/negative effect to the survey plan thoroughness. From the 99 grades attributed by each researcher

(11 different arrangements of item-attribute for each one of the nine survey plans), we identified agreement in 70 concepts, suggesting strong consistence of opinion between them. Once the evaluation was predominantly subjective, we decided to preserve the difference of opinions between researchers, composing the following formula to calculate each survey plan score.

$$Score(i) = \frac{N_i(pos)}{22} \qquad (3),$$

where *Ni(pos)* is the amount of *positive evaluations* obtained for the plan *i* in both evaluations (22 evaluations in total, 11 from each researcher).

Table 5-2 presents the score obtained for each plan (in the columns, represented by constellation names), indicating when the researchers matched (M) or diverge (D) in their evaluations. One can see that researchers presented divergent opinion in less than the half of thoroughness items for each subject, expect by one. However, a third researcher analyzed this survey plan, partially agreeing with both previous evaluations driving us to preserve the concepts originally attributed. Both three researchers concluded that one subject (Draco) did not understand the task, delivering an invalid survey plan which led us to exclude its plan from the dataset.

Regarding the subjects performance, one can also observe that FWK group obtained more positive evaluations in general than ADH. At the same time, only one subject (FWK group) in the whole sample presented a score higher than 0.75. Since we observed that ADH and FWK distributions of individual scores were normal (using Shapiro-Wilk) but not homoscedastic (Levene test), we applied the non-parametric *Mann-Whitney* test to compare the distributions of scores obtained by both groups. Once we did not observe significant difference between the distributions, we could not reject $H_0$1.

The score obtained by each group in each survey plan *characteristic* (arrangement of thoroughness item-attribute of trustworthiness) was calculated as the mean of the scores obtained by the group's subjects to the characteristic (CScore). One can observe in Table 5-3 that FWK CScores were higher or equal in most of cases than ADH CScores, although presenting a surprising lower performance in the *sampling frame* attributes. On the other hand, great differences could be observed in all recruitment process attributes. Since only the distribution of ADH scores was normal (Shapiro-Wilk), we applied the non-parametric *Wilcoxon signed-rank* test in order to analyze the significance of the differences observed between both distributions. As result, no significant difference were identified (p-value= 0.236), not allowing us to reject $H_0$2.

Table 5-2. Researchers' evaluation and score reached by each subject.

Thoroughness item	**Attribute of trustworthiness**	**Effect**								
		Cygnus ADH	**Draco ADH**	**Pegasus ADH**	**Hydra ADH**	**UrsaMajor ADH**	**Phoenix FWK**	**Andromeda FWK**	**LeoMinor FWK**	**Monoceros FWK**
Target Audience/ Unit of Analysis	Truth value	M(-)	D	M(+)	D	M(-)	M(-)	D	D	M(+)
	Consistency	D	M(-)	M(-)	M(+)	M(-)	D	D	M(+)	M(+)
Participants attributes	Truth value	M(+)	M(-)	M(+)	M(-)	D	D	M(-)	M(+)	D
	Consistency	M(-)	M(-)	M(+)	M(-)	D	M(-)	M(-)	D	M(+)
Sampling frame	Applicability	M(-)	M(-)	D	M(+)	D	M(-)	D	D	D
	Consistency	D	M(-)	M(+)	M(+)	D	M(-)	D	M(-)	M(+)
Sampling design	Applicability	M(-)	M(-)	M(+)	M(+)	D	M(-)	M(+)	D	M(+)
	Consistency	M(-)	M(-)	M(+)	M(+)	D	D	M(+)	D	M(+)
Recruitment Process	Consistency	D	M(-)	M(-)	D	D	M(-)	M(+)	D	M(+)
	Consistency	M(-)	M(-)	D	M(+)	M(-)	M(-)	M(+)	M(+)	M(+)
	Consistency	M(-)	M(-)	M(-)	M(-)	M(-)	M(+)	M(-)	M(+)	M(+)
Total (+)		5	1	14	14	7	5	12	14	20
Score		0.2273	0.0455	0.6364	0.6364	0.3182	0.2273	0.5455	0.6364	

Table 5-3. Score reached by each group for each characteristic evaluated.

Thoroughness item	Attribute of trustworthiness	ADH	FWK	Difference
Target Audience/ Unit of Analysis	Truth value	0.375	0.500	-0.125
	Consistency	0.375	0.750	-0.375
Participants attributes	Truth value	0.625	0.500	0.125
	Consistency	0.375	0.375	0.000
Sampling frame	Applicability	0.500	0.375	0.125
	Consistency	0.750	0.375	0.375
Sampling design	Applicability	0.625	0.625	0.000
	Consistency	0.625	0.750	-0.125
Recruitment Process	Consistency	0.375	0.625	-0.250
	Consistency	0.375	0.750	-0.375
	Consistency	0.000	0.750	-0.750

5.2.4. Follow up Results and Analysis

Table 5-4 shows the distribution of responses to the closed questions (*Likert* scale, four levels) regarding the *acceptance* of the technology (usefulness, easiness of use, intention to use). As indicated in the table, some of these statements were designed to be answered by only one group. In order to support verifying the consistence of the responses, we introduced some redundant statements and some negative ones. Except by the statement *"I followed all conceptual framework activities and tasks"*, all other statements presented only positive results. Subjects indicated that the framework concepts, activities, tasks and recommendations are easy to use and useful. In addition, one can see a predominant agreement on recommending and using the conceptual framework in future surveys. Thus, hypothesis $H_0 3$ could be rejected and hypotheses $H_A 3.1$, $H_A 3.2$ and $H_A 3.3$ could be accepted. However, the division of opinions between "partially agree/disagree" and "totally agree/disagree" observed in most of *usefulness and easiness of use* statements indicates that we need to investigate opportunities for improving the framework acceptance.

The follow up questionnaires also included open questions. FWK subjects were asked about the perceived positive/negative aspects on using the framework, while ADH were asked to describe in which extent the framework could be helpful on performing their tasks. Suggestions for improving the framework were also asked for the subjects

from both groups. Three FWK subjects reported positive and negative aspects on using the conceptual framework. Two of them pointed out difficulties on following the conceptual framework due to their inexperience on planning surveys. One reported difficult on distinguishing *unit of observation* from *unit of analysis*, while another asked by examples, including a complete example of a survey plan. The positive aspects reported by subjects emphasize the framework usefulness (*"…framework concerns with aspects that if are not observed, they could hamper the survey execution…"*[14]) and easiness of use (*"…In general, the framework guide me in a very explanatory way…"*). In addition, all ADH subjects pointed out that the conceptual framework would help them on better performing their tasks, especially due to their inexperience on planning surveys and due to the perceived relevance of its recommendations.

Five subjects from both groups presented improvement suggestions, all of them asking for introducing partial (search string, unit of observation, unit of analysis) or complete examples regarding the framework use. One subject also suggested reorganizing the framework to gradually introduce the concepts as they are mentioned in the framework activities/tasks. In addition, two subjects (ADH) asked by support on selecting the sampling design. In general, responses to the open questions helped us strengthen the indication that subjects tend to perceive the conceptual framework as useful (H_A3.1) and easy to use (H_A3.2). However, it also allowed us identifying specific but relevant limitations.

5.2.4.1. Recommendations Relevance Analysis

Subjects from both groups were also asked to indicate in the follow-up questionnaire the perceived five *more relevant* and five *less relevant* conceptual framework recommendations. Since subjects evaluated the recommendations from different perspectives, FWK and ADH opinions were initially analyzed in separate. As a result, we observed that ADH subjects predominantly perceived as less relevant recommendations addressing the *recruitment strategy* (14 from all 20 negative evaluations) while FWK subjects presented better distributed positive/negative evaluations.

[14] Translated from the Portuguese.

Table 5-4. Follow-up questionnaire answers (closed questions).

Acceptance Perspective	Statement	Totally disagree	Partially disagree	Partially agree	Totally agree
Easiness of use	All framework concepts are clearly described (Both)	0	0	4	4
	I understood all framework concepts (Both)	0	0	5	3
	I understood all framework tasks and activities (Both)	0	0	4	4
	I understood all framework recommendations (Both)	0	0	4	4
Usefulness	All framework recommendations are relevant (ADH)	0	0	0	4
	All framework recommendations are not relevant (FWK)	2	2	0	0
	The use of the framework was not relevant to understand better the planning of surveys in SE (FWK)	3	1	0	0
	The reading of the framework was not relevant to understand better the planning of surveys in SE (ADH)	3	1	0	0
	Using the framework was relevant to perform my task (FWK)	0	0	2	2
	I followed all framework activities and tasks (FWK)	0	*1*	1	2
	I followed all recommendations described in the framework (FWK)	0	0	1	3
Intention to use	I would use the framework to plan future surveys (Both)	0	0	1	7
	I would not recommend the framework to another researchers (FWK)	4	0	0	0

By aggregating the classifications provided by all the eight subjects in a single distribution we observed that nine recommendations presented negative evaluations in the third quartile (three or more negative evaluations): R06, R10, R11, R13, R18, R20, R23, R25 and R26. In the same way, eight recommendations could be classified as more relevant, being positively evaluated by two or more subjects (third quartile): R01, R02, R05, R07, R12, R16, R19 and R21. Finally, only two recommendations (R08 and R17) were not classified as more/less relevant by any subject. Such results suggest that there are recommendations perceived as more/less relevant than others, allowing us to reject $H_0$4 and accepting H_A4.1. Such unbalanced distributions suggest the set of recommendations need to be improved.

5.2.5. Discussion

The presented results indicate the conceptual framework v2.2 was accepted by the subjects but it was not helpful to improve the survey plans thoroughness, especially due to the low quality of the reported sampling frames. Analyzing the four sampling frames proposed by the conceptual framework users (FWK), we observed that two subjects reported vague sampling frames (*"requirement analysts working in software projects", "software factories working with software requirements"*), justifying the negative concepts received. In addition, another FWK subject apparently has confounded *sampling frame* with *sampling design*, even describing the sample size.

Thus, we decided to investigate how such results were influenced by issues in the conceptual framework documentation. In a big picture, the conceptual framework v2.2 (DE MELLO and TRAVASSOS 2015) introduces the new concepts of *source of population*, *search unit* and *population search plan* in order to support the composition of adequate sampling frames when they are not explicitly available. In this sense, the definition of source of population states that: "*…If a source of population can be considered 'valid' in a specific research context, it can be concluded that sampling frames can be established from it for the same research context.*" Then, it is established in the documentation that *Select the source of population (SP)* and *Design the population search plan (PS)* tasks should be executed before performing the activity *Design Sampling Strategy (SA)*, in which the sampling frame is established (task SA01). However, we also observed the absence of an explicit reference in the documentation associating the population search plan execution and the sampling frame composition. Indeed, only an implicit reference about it was found in SA01 ("*The sampling frame can be composed by all the search units available in the survey population, resulted from the execution of the previous activities*").

The issues observed in the documentation led us to suspect that the establishment of sampling frames by some subjects was probably performed without taking into account previous activities results, especially the design of the population search plan. In addition, once the concept *source of population* was not explicitly requested in the experimental task, it can be even possible that subjects bypassed activities and tasks addressed to such concept (indeed, two conceptual framework users declared not have followed all framework activities and tasks). In this sense, we identified the following opportunities for improving the conceptual framework documentation:

- To clear state the *population search plan* shall not be only *designed* (when needed) but also *executed* to retrieve the population available, supporting the sampling frame composition;
- To introduce examples addressing the involved concepts, and;
- To clear state the establishment of a *source of population* and a *population search plan* is not needed when there is an adequate *sampling frame* available.

The low scores obtained for *participants' attributes* in both groups can be explained due to the non-relevance of the control attributes provided by most of subjects. Extraneous conditions such as *"to be working at least by eight years as system analysis"*, *"to have between 25 and 35 years old"*, *"to have at least three years of experience in Software Engineering"* were reported, which can be explained by subjects inexperience with survey research and, more specifically, by the low experience in requirements engineering of the two FWK subjects presenting worst results in the tasks. We also observed that most of subjects from both groups reported only *control attributes* (attributes used to restrict the audience), although tasks TA03 (*Establish the unit of analysis attributes*) and TA04 (*Establish the unit of observation attributes*) state to first establish a set of control attributes and then enumerating the other attributes that should be collected, defining how to measure each one. Thus, no relevant opportunity on improving the framework addressing the survey units characterization was identified. Regarding the low scores of *truth value* obtained to the set of *target audience* and *unit of analysis* reported by FWK subjects, we observed again that two subjects' low experience in requirements engineering influenced in the negative concepts received, since they provided irrelevant/incomplete target audiences to the context of the survey, such as *"business analysts"* and *"software teams of medium/high complexity projects"*.

Thus, we can assume the low results obtained in the experiment execution were partially due to issues in the documentation, partially due to subjects' inexperience in the survey topic and also probably due to some conceptual framework users did not have follow one or more framework tasks.

5.2.6. Threats to Validity

Both researchers involved in the survey plans evaluation are from the same research group of the researcher that built the conceptual framework, which can be considered an internal threat to validity. These two researchers also lectured the Experimental Software Engineering course. To avoid bias in their evaluations, they received the edited survey plans without any subjects' personal identifications. Other internal threat to validity is due to the fact that one of the researchers involved in the survey plans evaluation also participated in the original survey execution (the one used as instrument) which could bias his opinion regarding the survey plans filled by the subjects. However, as already mentioned, the concepts provided by both researchers matched in most cases (more than 70%). The lack of control in the experiment execution is another important threat. Indeed, limitations of time in the ESE course led us to ask subjects to perform their tasks out of a controlled environment, such as the classroom. Thus, we cannot assure that they didn't share any information although no suspicious similarity between survey plans' items was observed.

Regarding construct validity, one can see that the performed experimental task performed was little different from survey planning in the practice. First, typically more than one researcher is involved in a survey planning; second, subjects were invited to complete a survey plan instead of designing a new one. Such experimental design was chosen taking into account the small sample size available and the concern with driving subjects' effort to the survey planning activities hypothetically supported by the conceptual framework. As external threat to validity, one can see that the small sample involved in the study is inexperienced in survey research, as well as they had declared never conducting any kind of study. This scenario limited our observation to a restricted context far to be representative from the state of practice, but interesting enough to allow us to evaluate and improve the conceptual framework.

5.3. Focus Group

Focus group has been shown a helpful research method for supporting *in depth* evaluation of SE technologies (HÖST, ORUČEVIĆ-ALAGIĆ and RUNESON, 2012; RODRÍGUEZ *et al.*, 2013, MOE *et al.*, 2014). Through focus group sessions, individuals are typically invited by a *moderator* to discuss in group and between groups different points of view regarding a specific topic. Meanwhile, one or more researchers take notes of the discussion, supporting posterior analysis. The discussion of benefits and drawbacks on using FG to support three different researches from the ESE Group

(including this research) is presented in (FRANÇA *et al.*, 2015). We planned a focus group session to investigate in which extent the conceptual framework v2.2 recommendations contribute or not to perform the survey planning task applied in the controlled experiment. From the 27 recommendations, a subset of eight recommendations was considered in the scope of the focus group, based on the following criteria due to session time constraints:

- Recommendations more frequently classified as less relevant by the subjects in the follow up questionnaire, related to survey plan items worst evaluated in the experiment (R06, R10, R11, R13 and R18);
- Recommendations not classified by any subject as more/less relevant in the follow up questionnaire (R08 and R17);
- The single recommendation classified six times as more relevant by the subjects, to be used as control (R12).

Following sections describes the focus group plan and execution, reporting its results and the opportunities identified for improving the recommendations.

5.3.1. The Plan

The focus group planning activities involved five ESE Group researchers, four of them with previous experience on applying such technique. Taking into account the different perspectives followed by FWK and ADH subjects on performing their tasks and the distinct results observed on evaluating the conceptual framework v2.2 recommendations, we decided to preserve the same groups' composition. Therefore, in the focus group context, FWK subjects composed the *users group*, since they effectively applied the conceptual framework, while ADH subjects composed the *readers group*, since they only read the conceptual framework documentation. Besides the *moderator*, two researchers should act as *scribes*, taking notes of the discussions (one for each group), and a third researcher should act as *observer*, reporting behaviors, attitudes and reactions expressed by all other participants. Figure 5-2 presents the planned physical distribution of the participants in the classroom.

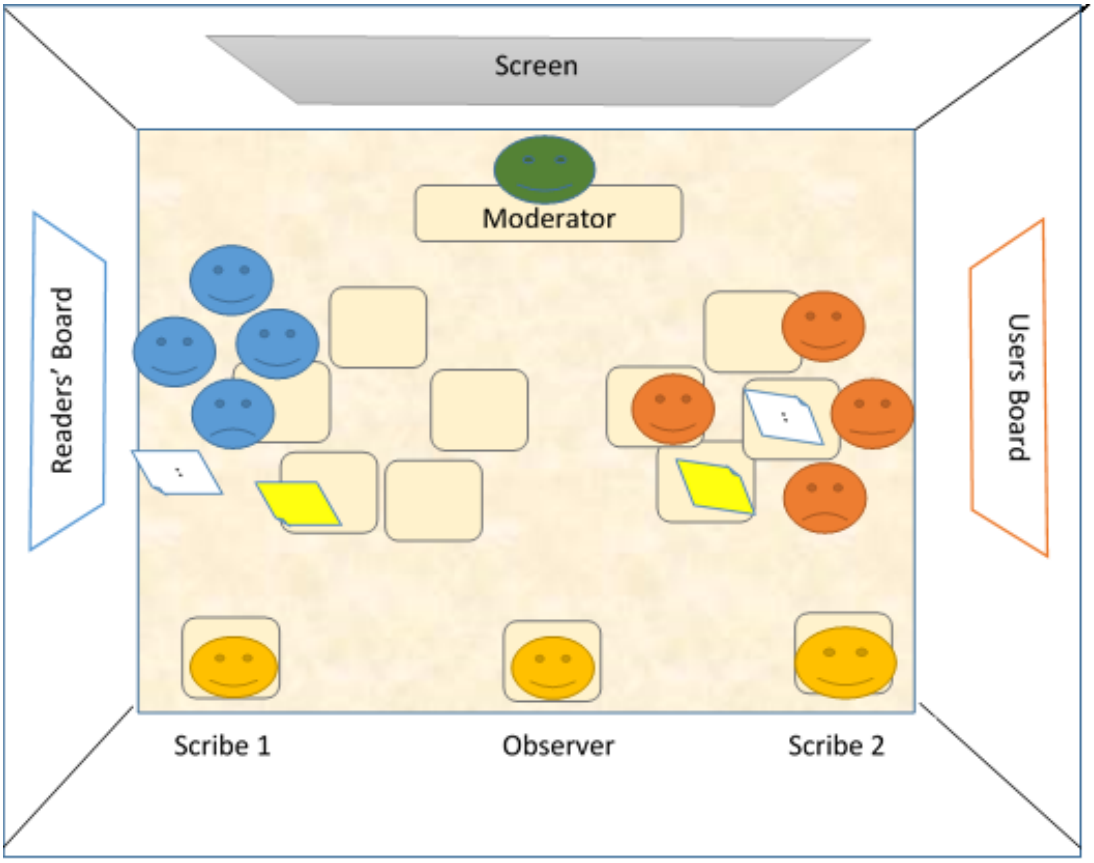


Figure 5-2. The planned physical distribution of the participants and instruments.

The focus group session was planned to be executed in two sequential phases. In the first phase, the moderator should distribute subjects in the room and present the activity dynamics. Then, the moderator should distribute *pens*, *post-its*, and a *printed copy* of the framework documentation (concepts in yellow and activities in white). Our intention on printing framework components in different colors was to observe in which extent both groups will consult different parts of the documentation. In the second phase, the following steps should be repeated until all eight recommendations have been discussed, observing the sequence (R08, R10, R11, R12, R13, R17, R18, and R06):

1. Moderator presents a recommendation.
2. Moderator asks subjects to discuss in group why the presented recommendation contributes/ does not contribute for planning surveys, writing their arguments in post-its and fixing them in the group board (Table 5-3).
3. Moderator invites Readers to present their favorable/unfavorable arguments regarding the recommendation.
4. Moderator invites Users to argue on Readers' opinion based on their annotations, starting the discussion.

CONTRIBUTES		DOES NOT CONTRIBUTE
R08		R08
R10		R10
R11		R11
R12		R12
R13		R13
R17		R17
R18		R18
R06		R06

Figure 5-3. Board designed to each group fix their arguments, written in post-its

5.3.2. Execution

The focus group session was conducted during an ESE course class (Figure 5-4). All feasibility study's subjects participated. Two other students that didn't participate in such study but had attended to the class were distributed in the groups. We originally planned to allocate two hours from an ESE course for the study execution but only one hour and half was available. Thus, we decided to change the second phase *on the fly*, preserving the planned steps 1 and 2 to each recommendation but changing steps 3 and 4 to include all recommendations addressed to a specific survey item in a single session. Thus, all eight recommendations were discussed in three distinct sections: a first one discussing recommendations addressed to the composition of the *sampling frame* (R08, R10, R11, R12 and R13); a second one involving recommendations addressed to the *sampling design* (R17 and R18); a third section devoted to the single recommendation related with *participants attributes* (R06).

At the end, the whole focus group meeting effectively took one hour and 53 minutes. Phase one took approximately 30 minutes and the time spent in phase two was approximately balanced between the sections taking into account the number of recommendations evaluated in each section. All subjects participated in all sections, except one subject from the users' group that did not participated in the third session. Some subjects presented an unexpected difficult on interpreting English expressions which could influenced them on misinterpreting some recommendations[15]. However, scribes were aware and helped translating recommendations when needed.

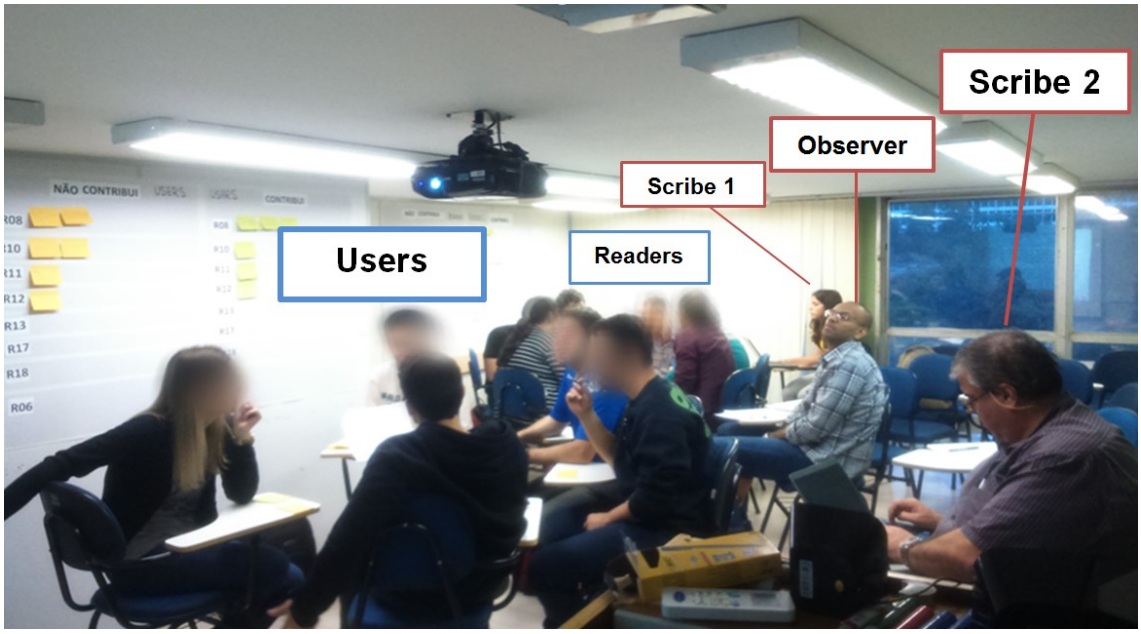


Figure 5-4. Distribution of the focus group participants in the room. Picture taken by the moderator.

5.3.3. Results and Analysis

Except by R08, all recommendations have received favorable and unfavorable arguments, totalizing 20 arguments reported by readers and 20 arguments reported by users in the post-its (Figure 5-5 and Figure 5-6, respectively). However, we identified that some presented arguments were, in fact, statements/speculations regarding the recommendations' usefulness/validity. Although the moderator had explained (and reminded during the sessions) that the focus group's goal was to discuss *why* each recommendation contributes/does not contribute for the survey planning task, the group discussions were often deviated from its goal. For instance, subjects stated (without arguments) that: *"using SE conferences as source of population is not a good idea"*[16]; *"LinkedIn is not a useful source of population"*; *"survey pilots should be avoided when the source of population in hard to be accessed"*. In addition, subjects also reported improvement suggestions since the first section, which was stimulated by the moderator again in the other sections. We coded the arguments provided by the subjects and the notes from scribes/observer to clarify what are perceptions/feelings related to each recommendation.

[16] Translated from Portuguese.

Figure 5-5. Readers' panel after the execution of the focus group (*Não contribui= Does not contribute; Contribui= Contributes*).

Figure 5-6. Users' panel after the execution of the focus group.

Following subsections presents data analysis performed over the coded arguments and suggestions provided to each recommendation. Highlights from the discussions are also presented. For the better comfort of the reader, we reproduce in each subsection the corresponding recommendation description.

5.3.3.1. Recommendation R08

''If the target audience is limited to SE researchers (or subsets from them), valuable sources of population includes digital libraries such as SCOPUS and IEEE while the search unit can be the papers available in such sources. Professional social networks such as ResearchGate and Academia.edu can be also useful for searching SE researchers."

1. *Favorable Arguments:* practical (R, U)
2. *Unfavorable Arguments:* intuitive (U), restrictive (U)
3. *Suggestions:* to introduce other examples of sources, such as Lattes platform (Brazilian official database of researchers) and papers authors (U).
4. *Discussion Highlights:* strongly influenced by Monoceros, users claim that R08 is intuitive which is strongly rejected by readers. Orion argues that concept of intuitiveness is relative and ironically asks: *"What is not intuitive after you know?"*
5. *Analysis:* The structured review suggests that R08 is not intuitive since only few works following such practices were found. Additional examples (worldwide) of sources can be included. Authors referenced in papers are not sources of population, they are the subjects but the set of papers included in SLR can be useful to identify subjects.
6. *Orientation:* To *preserve* R08, creating an additional recommendation specifically driven to reusing SLR results.

5.3.3.2. Recommendation R10

"*The professional social network LinkedIn has been showed a useful source to access representative samples composed by SE professionals through their groups of interest."*

1. *Favorable Arguments:* none
2. *Unfavorable Arguments:* intuitive (U), incomplete (U)
3. *Suggestions:* To describe how to perform searches in LinkedIn (U).
4. *Discussion Highlights:* Users (Monoceros) insists on using intuitiveness as an unfavorable argument which is continuously refuted by the readers. One scribe intervened pointing out that recent investigations suggest that using LinkedIn as source of population is not intuitive. Chameleon rhetoric made an inhibitive effect over Readers. Subjects devoted many time discussing if LinkedIn is valid or not as source of population, which was out of scope of the focus group.

5. *Analysis:* LinkedIn is barely observed as source of population in the specialized literature, therefore it is not intuitive. Describing how to use the LinkedIn could be exhaustive and risky due to the lack of control on how the LinkedIn could make available/restrict its search resources in the future.
6. *Orientation:* To *evolve* R10, driving the recommendation to professional social networks in general and other types of source of population mentioned in R11.

5.3.3.3. Recommendation R11

"Other types of possible Source of population includes, but are not limited to: open discussion groups, SE conferences, project repositories and freelancing tools."

1. *Favorable Arguments:* essential (U), opportune (R)
2. *Unfavorable Arguments:* incomplete (U)
3. *Suggestions:* To group R11 with R08 and R10 (R).
4. *Discussion Highlights:* Users asked *"What is a freelancing tool?"*. Users also asked how to use each source. Again, subjects spent many time discussing if LinkedIn is valid or not as source of population.
5. *Analysis:* Limitations on using SE conferences should be explained.
6. *Orientation:* To *evolve* R11, addressing it specifically to SE conferences and explaining the limitations on using such type of source.

5.3.3.4. Recommendation R12

"Before taking your decision, invest efforts on simulating the use of each source and its search unit, since some technical limitations may be not explicit in a first glance. If you are not sure in which Source of population should be selected, consider running a pilot trial using each one to support your decision."

1. *Favorable Arguments:* Support validation (R)
2. *Unfavorable Arguments:* Unclear (U)
3. *Suggestions:* none.
4. *Discussion Highlights:* Apparently, the sense in which the terms "simulation" and "pilot" are used was not reached by subjects from both groups. Again, both groups spent their time discussing the validity of the recommendation.
5. *Analysis:* The beginning of the recommendation should be rewritten; the term simulating should be avoided (maybe substituted by evaluating) and the use of the term pilot should be better contextualized, in order to not be confounded with the survey pilot.

6. *Orientation:* To *divide* R12 into two recommendations, addressing one to evaluating sources and another to conducting pilots.

5.3.3.5. Recommendation R13

"*Report any special condition needed to use the selected Source of population. For instance, a LinkedIn "Premium" account was needed to make feasible some group analysis identify members in common (overlapping) between different groups of interest."*

1. *Favorable Arguments:* Supports reuse (U, R)
2. *Unfavorable Arguments:* Out of context (U)
3. *Suggestions:* none
4. *Discussion Highlights:* Some difficult on interpreting the text (English) were presented by some subjects. Users started to ask themselves if the framework is LinkedIn-oriented.
5. *Analysis:* The framework support planning activities. Thus, recommendations addressed to survey plan reusability are pertinent.
6. *Orientation:* To *preserve* R13.

5.3.3.6. Recommendation R17

"If there is enough information in the units' attributes that clearly allows you to extract homogenous subpopulations for supporting clustered sampling, evaluate the feasibility of applying stratified sampling."

1. *Favorable Arguments:* Preventive (U)
2. *Unfavorable Arguments:* Incomprehensible (R), Unclear (U)
3. *Suggestions:* none.
4. *Discussion Highlights:* Most of subjects did not understand the recommendation, even after a scribe fully translated it. Each group tried to "spy" the other group, paying attention in what they were talking.
5. *Analysis:* It was identified the missing of a "NOT" in the beginning of the recommendation ("If there is NOT enough…"). The recommendation is pretty specific and barely contextualized since sampling designs are not introduced in the framework.
6. *Orientation:* To *remove* R17. To introduce the more common sampling designs in the framework through designing a set of tasks and recommendations addressed to them.

5.3.3.7. Recommendation R18

"*When calculating the sample(s) size(s), have in mind that participation rates in voluntary SE surveys over random samples tend to be small (less than 10% of the sample size).*"

1. *Favorable Arguments:* Preventive (U, R)
2. *Unfavorable Arguments:* Intuitive (R)
3. *Suggestions:* none.
4. *Discussion Highlights:* Readers and moderator started to express fatigue. Readers were divided on discussing the intuitiveness of the recommendation. Users asked themselves in which context the small participation rates were observed.
5. *Analysis:* Insufficient argument was provided to classify the recommendation as intuitive. In the other hand, one can see the recommendation is not a rule of thumb and should be better contextualized.
6. *Orientation:* To *evolve* R18, letting clearer that it is based on observing large scale surveys using non-convenience samples composed by practitioners.

5.3.3.8. Recommendation R06

"*When possible, follow already established standards to support the measurement of attributes. Standards can be especially helpful in the case of attributes measured through scales and nominal variables.*"

1. *Favorable Arguments:* Realistic (R), Preventive (U)
2. *Unfavorable Arguments:* Incomplete (R, U)
3. *Suggestions:* To introduce examples (R, U).
4. *Discussion Highlight:* Although their fatigue, all subjects from both groups have participated and interacted. Readers' members were initially in doubt about which is a "unit of observation". Both groups reached a consensus on the need to providing examples.
5. *Analysis:* Examples should be added.
6. *Orientation:* To *preserve* the recommendation, introducing examples.

5.3.4. Discussion

The focus group ran without relevant interruptions. Moderator and scribes intervened in the discussions a couple of times to stimulating the participation and explaining methodological issues but it was not observed bias in such interventions.

Mood and behavior changes were barely observed and all researchers involved concluded that focus group session was a little tepid. We observed moderate synergy among participants, being more intense on discussions *within* groups than *between* groups. Indeed, a couple of participants from each group were significantly more active than others in the discussions between groups, prevailing constantly their opinions. Since participants were not playing roles, they were free to present their actual opinions (positive/ negative) regardless whether such opinions are contradictory or consonant with others. In this sense, notes taken by the observer reported rare situations in which participants expressed behaviors that could put in check their arguments. In most of cases, groups presented different arguments and preserved their positions until the end of the discussions, helping us to better understanding their reasons.

The follow up questionnaire was used to ask subjects to classify the five more/less relevant recommendations while focus group asked to explain why each recommendation contributes or does not contribute to perform the survey plan tasks. Thus, our expectation was that recommendations frequently classified as *less relevant* would be associated with *unfavorable* arguments while recommendations frequently classified as *more relevant* would be associated with *favorable* arguments. However, the comparison between the incidence of favorable/unfavorable arguments and the relevance reported in the follow up questionnaire to each recommendation allowed us to observe many deviations from this expectation. The focus group findings suggest its dynamics probably influenced *users* on changing their opinion regarding many recommendations, while *readers*' opinion was perceived as more consistent between both activities. One possible explanation for this relies on the fact that only users could share their individual experiences on using the framework, which could lead them on reflecting in more depth about the arguments to be provided and eventually changing their points of view.

5.3.5. Threats to validity

The focus group moderator was also the researcher who conceived the framework, which is a threat to internal validity, as it may have influenced the opinions of the participants. However, the other three involved researchers observed that his behavior during the focus group session was predominantly neutral. The unexpected limitation in English language identified in some subjects is another important threat that interactive and exploratory nature of focus group allowed us to observe. In fact, it probably also influenced their negative results in the feasibility study. However, all English-related issues presented by the subjects during the focus group were clarified.

In this context, it is important to highlight that although the framework documentation was provided in English all activities were conducted in Portuguese (native language of all participants).

The often deviation from the focus group discussion could be explained by the subjects' and specially moderator's inexperience in focus groups. The lack of subjects' background on planning surveys should be also taken into account. We tried to mitigate such limitations through orienting subjects to keep their focus on exploring the relevance of each recommendation from the point of view of their experience performing the experimental task. We also highlight that two students that did not participated in the feasibility study participated in the focus group, being such activity their first contact with the conceptual framework.

5.4. Conclusion

This Chapter presented the empirical studies conducted for evaluating the conceptual framework v2.2. First, a feasibility study was conducted in vitro to characterize the acceptance of the proposed technology and its contributions to the thoroughness of the surveys plans. Although it was evidenced the feasibility of the proposed technology, we could not identify a comprehensive contribution of the conceptual framework v2.2 to improve the thoroughness of the survey plans. In addition, we evidenced that some framework recommendations were perceived as more relevant than the others. In this sense the focus group session conducted allowed the emerging of important opportunities for evolving the set of the conceptual framework recommendations. As a result a third new version of the conceptual framework presented in Chapter 6 was depicted.

6 The Conceptual Framework v3

In this chapter, we present the third version of the conceptual framework evolved from the empirical evaluations performed over its previous version.

6.1. Introduction

We developed the third (and current) version of the conceptual framework based on the findings of the empirical studies presented in the Chapter 5. Table 6-1 presents the conceptual framework v3 concepts and their respective properties, indicating whether the instantiation of each concept is *mandatory* (M) or *conditional* (C) and whether their properties should be *mandatory* (M) or *optional* (O). The content of this chapter is heavily based in (DE MELLO and TRAVASSOS, 2016).

Minor changes were made in the conceptual framework concepts. Since the unique possible type of *unit of observation* in opinion surveys is the *individual*, we renamed such concept to *study subject* in order to avoid mixing it with *unit of analysis*. Since *inclusion criteria* concept can be diverse from the opposite of *exclusion criteria*, we included such property to the *population search plan*. We take out *sampling frame* from the *sampling strategy* concept, since a single sampling frame can be reused to support different sampling strategies. On the other hand, we introduced *confidence interval*, *confidence level* and *sample size* as *sampling strategy* properties. Finally, we decided to preserve the idea of designing *characterization questions* without the concern on designing the instrument (typically the *characterization questionnaire*), task considered out of the conceptual framework scope.

While few changes were made in the conceptual framework concepts, we significantly evolved its activities, tasks and recommendations. In general, we identified the survey planning process proposed in the conceptual framework v2.2 was not clear about activities/tasks conditionality. For instance, if the research has an adequate sampling frame available (i.e., composed by a representative population from the point of view of the survey objective), there is no need to look for sources of population and to design a population search plan. Other important changes are addressed to include guidelines to support the adoption of specific sampling designs and the distinction between the population search plan design and its execution. Figure 6-1 present the

survey planning process adapted with the activities designed to support the use of the conceptual framework concepts.

Table 6-1. Concepts and properties from the third version of the conceptual framework.

Main Concept	Mandatory/ Conditional	Properties	Mandatory/ Optional
Target Audience	M	-	
Subject (unit of observation)	M	Attributes	M
Unit of analysis	M	Type of Entity	M
		Attributes	M
Sampling Frame	M	-	
Source of Population	C	Description	M
		Search Unit	M
Population Search Plan	C	Search String	O
		Search Algorithm	M
		Inclusion Criteria (new)	O
		Exclusion Criteria	O
Sampling Strategy	M	Sampling Design	M
		Sample Size (new)	M
		Confidence Interval (new)	M
		Confidence Level (new)	M
Recruitment Strategy	M	Invitation Message	M
		Execution Estimated Time	O
		Period Available	M
		Reminding Method	O
		Reward Method	O
Subject Characterization Questions (new)	C	-	
Unit of Analysis Characterization Questions (new)	C	-	

Table 6-2 presents the instantiation of the conceptual framework v3 to support the same survey used as example in Chapter 4. Section 6.2 describes the conceptual framework v3 activities with its tasks and recommendations. Since minor changes were made in the conceptual framework concepts' we opted by presenting it in the Appendix B.

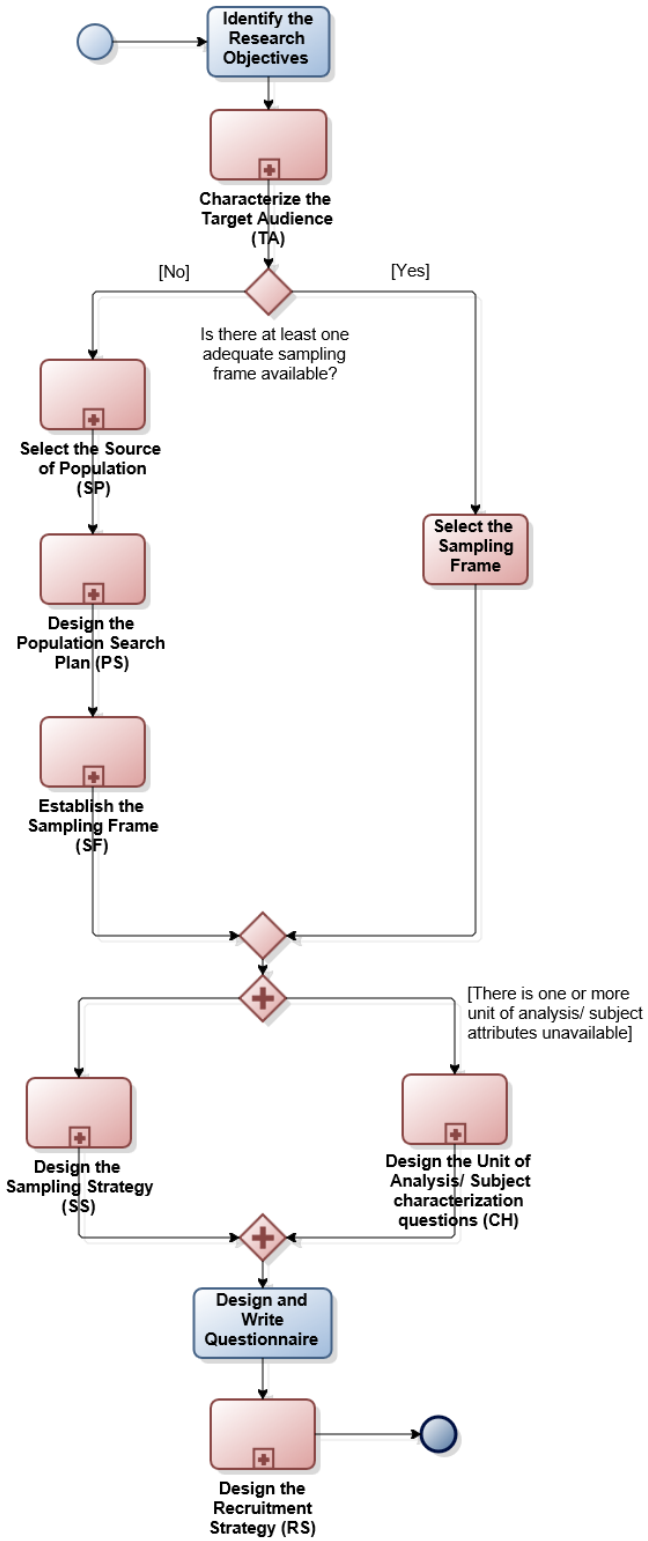


Figure 6-1. The survey process adapted by the new proposed version of the conceptual framework.

Table 6-2. Example of instantiating the conceptual framework v3 to support a survey on SLR usage in SE.

Concept	**Property**	**Properties**
Target Audience	-	Brazilian SE Research Groups
Unit of analysis	Type of Entity	Research Group
	Attributes	Control attributes: *location= "Brazil" and research field= "SE"* Other attributes: *name, description, location (city), age, number of doctors, number of doctoral students and number of master students*
Subject	Attributes	Higher Academic Degree= *"PhD", "DSc" OR "Master"* Other attributes: *researcher name, time in the research group (years), experience with SE research, experience with SLR*
Source of Population	-	CNPq research group directory (http://dgp.cnpq.br/dgp/)
	Search Unit	Research Groups
Population Search Plan	Search String	*"software engineering"* OR *"engenharia de software"*[17]
	Search Algorithm	Apply each search string (between quotes) once in the field *"Search term"*, choosing the options *"Exact search"* and *"search by group".* Check to apply each search only to the following fields: *"group name", "research line name"* and *"keyword of the research line"*. Do not apply other filters available.[18]
	Inclusion Criteria	None
	Exclusion Criteria	To exclude all research group retrieved created only after 2010 To exclude any research group in which description do not clear mention the use of empirical (experimental) methods to support their research in SE
Sampling Frame	Description	The list of all research groups resulted from the execution of the Population Search Plan
Sampling Strategy	Sampling Design	Simple Random Sampling of the research groups available in the sampling frame, selecting all the subjects available in each group.
	Confidence Interval	5 points
	Confidence Level	95%
	Sample Size	*To be calculated based on the results of the population search plan*

[17] "Software Engineering" in Portuguese

[18] The actual names of the fields are in Portuguese

Recruitment Strategy	Description	Since the e-mail of the subjects is not available in the source, to send a standard e-mail to the leader from each research group, asking him/her redirect the survey invitation
	Invitation Message	*To be described*
	Estimated Execution Time	Not estimated
	Period Available	10th January 2016 - 24th January 2016 (15 days)
	Reminding Method	In the 7th day of survey execution, to re-send the invitation message only to the leaders of those research groups that no response was received
	Reward Method	None
Subject Characterization Questions	-	*"How long (years) you have been working with SE research?"*
Unit of Analysis Characterization Questions	-	None

6.2. Conceptual Framework v3- Activities

Following subsections introduce each conceptual framework v3 activity, describing its tasks and its specific recommendations.

6.2.1. Characterize the Target Audience (TA)

Since the survey research objectives were established, you are able to identify and characterize its target audience. Thus, you should perform the three following tasks, in the presented sequence. Table 6-2 describes their respective recommendations.

TA01. *Identify the target audience.*

R1. Based on the research objectives, answer the following question: *"Who can best provide you with the information you need?"* Thus, at this moment, try to avoid restricting the target audience based on factors such as its size or population availability.

TA02. *Characterize the unit of analysis.* Based on the survey target audience, Identify the entity from which survey data will be analyzed and establish its attributes.

R2. Establish first the set of control attributes and their respective values that will restrict the unit of analysis. Then, enumerate the other attributes that should be collected from each unit of analysis and define how to measure each one.

R3. It has been observed that the own subject (individual) is the more frequent entity used as unit of analysis in SE surveys, followed by *organizations*, *organizational units* and *project teams*.

R4. Individuals can be characterized in SE surveys through attributes such as: *experience in the research context, experience in SE, current professional role, location and higher academic degree*.

R5. Organizations can be characterized in SE surveys through attributes such as: *size (scale typically based in the number of employees), industry segment (software factory, avionics, finance, health, telecommunications, etc.), location and organization type (government, private company, university, etc.)*.

R6. Project teams can be characterized through attributes such as *project size; team size, client/product domain (avionics, finance, health, telecommunications, etc.) and physical distribution*.

R7. Taking into account the limitations of each data type involved, you can establish formulas to measure characteristics composed by two or more attributes. For instance, DE MELLO, DA SILVA and TRAVASSOS (2015) calculated the subject' experience level through a weighted mean between four distinct dimensions: highest academic degree (scale); number of years working with SE; number of years working with the research scope; number of SE projects already concluded (scale).

R8. When possible, look for already established standards to support the characterization of the survey unit of analysis/subjects. Standards can be especially helpful to provide scales and even nominal values. For instance, CMMI-DEV maturity level can be used to characterize software organizations regarding their maturity in software process, while one can use RUP roles to characterize subjects' current position.

TA03. *Establish the subject attributes.* This step should be performed whether the unit of analysis is not the own individual (the survey subject). Thus, recommendations R02, R04, R07 and R08 are applicable.

6.2.2. Select the Sampling Frame

If you are sure to have access to one or more suitable sampling frames that could provide you representative samples to support your survey, you need to select one to support your sampling activities. If don't (common issue in SE research) you should *select the source of population* (SP).

6.2.3. Select the Source of Population (SP)

This activity aims at selecting the source from which is expected to identify an accessible and representative population to support your survey. For this, you should perform the three following tasks, in the presented sequence. Table 9 describes their respective recommendations.

SP01 *Identify the candidates.* Look for candidates to source of population.

R9. Avoid the convenience on searching candidates, trying to answer: *"Where a representative population from the survey target audience or even all target audience is available?"*

R10. If the target audience is limited to SE researchers (or subsets from them), valuable candidates includes digital libraries such as SCOPUS being their papers the search unit. Social networks addressed to integrate academics such as *ResearchGate* and *Academia.edu* can be also useful in this context.

R11. Have in mind that results from previously conducted SLR regarding your research theme may provide representative populations composed by researchers (paper authors). Examples can be found in (DIAS-NETO and TRAVASSOS 2008; ABRANTES and TRAVASSOS 2013; CARVER *et al.*, 2013; GUZMÁN *et al.*, 2014)

R12. Look for catalogues provided by recognized institutes/ associations/ governments to retrieve relevant set of SE professionals/ organizations. For instance, SEI institute provides an open list of the organizations and organizational units certified in each CMMI-DEV level. FIPA provides information regarding Finland IT organizations and its professionals. CAPES, from the Brazilian government, provides a tool for accessing information regarding Brazilian research groups.

R13. Sources available in the web such as *discussion groups*, *projects repositories* and *worldwide professional social networks* can be helpful to identify representative populations composed by SE professionals. However, have in mind that such sources are not designed to support research and they can restrict at any moment the access to the content available, which can hamper your sampling activities. For instance, DE MELLO et al., successfully used LinkedIn groups of interest to support two surveys in 2013 (DE MELLO and TRAVASSOS 2013b; DE MELLO, DA SILVA and TRAVASSOS, 2015) and an on-line experiment in 2014 (DE MELLO, STOLEE and TRAVASSOS, 2015). However, in 2015 it was observed that recruitment of group members in this tool were significantly restricted.

R14. SE conferences concerned with the survey research objective may be good candidates. However, it is important to evaluate if you will have the opportunity to apply systematic steps for sampling and recruitment during/ after the conference.

SP02 Evaluate the Candidates. Apply the Essential Requirements (ER) to each candidate and discard any combination of source of population candidate/search unit that do not support all ER. Then, apply the desirable requirements (DR) to help you on reflecting about the advantages on using each one.

SP03 *Select the Source of Population.* Select a single source that you understand that could deliver a representative sample to support your research.

R15. Before taking your decision, invest efforts on exploring the source of population, searching by units of analysis/subjects, collecting their data and trying to certify that it will be possible to keep in touch with the subjects selected.

R16. Report in the survey plan any special condition needed to operate the selected source of population as planned. For instance, a LinkedIn "Premium" account was needed to make feasible the analysis of the similarities between the groups of interest selected (DE MELLO et al., 2015).

6.2.4. Design the Population Search Plan (PS)

This activity aims at supporting the extraction of the survey population from the selected source of population. For this, you should perform the following tasks. Table 10 describes their respective recommendations:

PS01. *Design the search algorithm.* A search algorithm must be designed to describe how the population will be searched in the selected source of population.

R17. In order to support the reuse of the survey plan, the search algorithm should describe any particularities and restrictions on manipulating the source of population. For instance, if the source of population is provided by a Web application, it is important to describe how to access and apply the search unit (parameters, option, menus). Have in mind that such resources may change in the future.

PS02. *Design the search string.* The use of search strings can be helpful on filtering the suitable search units to the survey context. Depending on the specialization of the source of population and on the search resources available, even complex strings may be needed, using logical operators as commonly used in systematic literature reviews (SLR).

R18. Consider consulting the specialized literature (especially standards such as IEEE vocabulary for Software Engineering) and/or specialists for identifying a wide range of relevant and similar expressions for composing the survey search string. If a SLR was previously performed in the context of your research, consider reusing its own search strings and data provided by its results.

PS03. *Design the inclusion/ exclusion criteria.* Design the set of criteria that will be used to qualitatively filter the results from the execution of the search algorithm.

R19. If the search unit allows retrieving groups of units of analysis instead of a single unit of analysis, identify relevant attributes to characterize each search result. Such attributes can be helpful to compose your inclusion/ exclusion criteria. For instance, DE MELLO, DA SILVA and TRAVASSOS, 2015 established LinkedIn group of interest as search unit and individual as unit of analysis. In this sense, the authors used group attributes such as its name and its description to support their decision on including/excluding each group of interest identified.

R20. Use inclusion/exclusion criteria only if it is actually necessary to reduce noise in your population. If you are sure that all search units retrieved from the source of population will be valid, avoid introducing selection bias in the exclusion/inclusion criteria.

R21. Have in mind that exclusion criteria is not necessarily the opposite of inclusion criteria. You can use both to support your filtering. However, avoid composing a set of inconsistent or even redundant criteria.

R22. Evaluate if the inclusion/exclusion criteria are composed by one or more conditions that can be automatically verified. If so, consider to insert them into the search algorithm/search string.

6.2.5. Establish the Sampling Frame (SF)

This activity aims at supporting the establishment of the survey sampling frame based on the results from the execution of the Population Search Plan. You should perform the following tasks in the presented sequence. Table 11 describes their respective recommendations.

SF01. *Execute the Population Search Plan.* Take note about when (date/time) the search plan was executed and save all the results obtained in each step.

SF02. *Evaluate the Results.* Evaluate if the population search plan execution provided a satisfactory result, i.e., a representative population to support the survey execution. If don't, you can review your population search plan or even select another source of population

R23. Do not be concerned only with the population size retrieved but also its heterogeneity. Have in mind the focus on providing target audience representativeness.

SF03. *Establish the sampling frame*.

R24. Take preference to select all the results retrieved in the population search plan execution. However, if it will imply in prohibitive operational efforts, investigate alternatives that could preserve the population representativeness. For instance, DE MELLO, DA SILVA and TRAVASSOS (2015) identified 64 groups of interest after executing the population search plan. However, since it was observed that 90% of the subjects was member from the 19 more populous groups of interest available, the researchers decided to sample only from these groups.

6.2.6. Design the Sampling Strategy (SS)

This activity aims at supporting how a sample will be extracted from the sampling frame, supported by the following tasks. Figure 6-2 presents how these tasks are organized and Table 12 describes their respective recommendations.

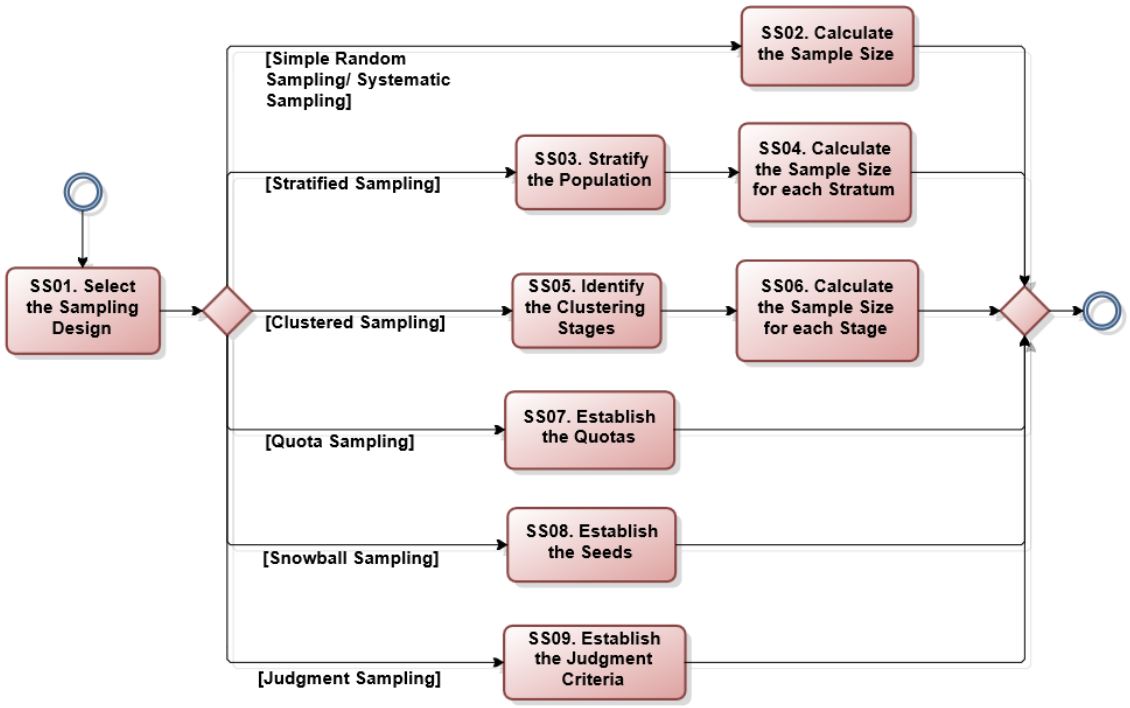


Figure 6-2. Tasks for designing the Sampling Strategy.

SS01. *Select the sampling design.* A single sampling design should be selected. Statistics literature presents the more common probabilistic sampling designs and formulas for calculating survey sample size (THOMPSON, 2012).

R25. If you successfully established a sampling frame composed by a representative population, you should apply a probabilistic sampling design. Otherwise, you will waste the opportunity to generalize the results obtained and to aggregate to future survey re-executions using the same population.

R26. If you selected a probabilistic sampling design, avoid introducing any non-probabilistic step on the sampling activities. For instance, in case of subjects being selected from a sample composed by a set of SE organizations, it is important to avoid the introduction the bias of organization representatives on selecting the subjects.

SS02. *Calculate the Sample Size* (Simple Random Sampling/ systematic sampling). Apply Simple Random Sampling for each subpopulation, using the same confidence level and interval.

R27. When calculating the sample size, have in mind that participation rates in voluntary SE surveys over random samples tend to be small. In large-scale surveys based on random samples composed by SE practitioners, it has been observed participations rate lower than 10%.

SS03. *Stratify the Population* (Stratified Sampling).

R28. When choosing stratified sampling, be sure that you can draw distinct subpopulations that will be relevant to support your analysis.

R29. Have in mind that no unit of analysis from the survey population can be let out from a stratum and all strata must be mutually exclusive, i.e. a single element cannot be found in more than one stratum.

SS04. *Calculate the Sample Size for Each Stratum* (Stratified Sampling). Apply Simple Random Sampling for each subpopulation, using the same confidence level and confidence interval to each one.

SS05. *Identify the Clustering Stages* (Clustered Sampling).

R30. Results obtained by clustered sampling are statically weaker than those obtained through simple random sampling. If you do not have technical restrictions to access and retrieve data from any part of your survey population, avoid this sampling design.

R31. Clusters should be naturally observed in the population for supporting clustered sampling. If you cannot assure there are similar unit of analysis in the population, clustering them probably will hamper the validity of the survey results.

SS06. *Calculate the Sample Size for Each Stage* (Clustered Sampling).

SS07. *Establish the Quotas* (Quota Sampling). Identify the criteria that you will use distribute the population in mutually exclusive subsets and how many subjects you will sample from each one.

SS08. *Establish the Seeds* (Snowball Sampling). Identify the subjects that will compose the first wave of the snowball. They should indicate additional subjects to be then recruited. Alternatively, researchers can choose to cyclically asking by new subjects establishing an arbitrary stop criterion.

SS09. *Establish the Judgment Criteria* (Judgment Sampling). Establish criteria for selecting the survey subjects from a sampling frame. For instance, you can decide to sample SE researchers in which curriculum presents your research theme as topic of interest.

6.2.7. Design the Recruitment Strategy (RS)

This activity aims at establishing how to recruit the sample for participating in the survey. You should perform the following tasks in the presented sequence. Table 13 describes their respective recommendations.

RS01. *Establish the invitation method.* Considering the individuals' contact data available, establish how each sample subject will be invited to answer the survey.

R32. Samples in SE surveys are often invited through e-mails. In such case, take preference to send individual and personalized messages using a standard message as a template. When, possible, activate e-mail read notification.

R33. In the case of the survey samples are distributed through discussion groups/ groups of interest in the Web, avoid posting a generic forum invitation message, except if you have control on whose and how many members will read it.

R34. Depending on the source of population used, no contact data may be available. Alternatively, professional social networks may allow researchers to keep in touch with the subjects using internal message resources. However, have in mind that social networks frequently restrict the use of such resources in large scale.

RS02. *Parameterize the recruitment.* Establish values to the recruitment parameters presented in the concept "Recruitment Strategy".

R35. A finite and not so long period to answer the survey allows subjects to plan themselves about when they will answer the survey (scarcity).

R36. It is common in SE surveys to send a single remind message. Sending many reminds in a short term, especially when participation is voluntary, can bring a negative effect in the participation.

R37. It is expected that rewards can stimulate the participation in surveys, but such practice is still uncommon in SE research. Rewards can be offered only to the subjects who successfully completed the survey questionnaire or even to all subjects. In international surveys, it can be hard to establish a low-cost payment instrument for small rewards. Alternatively, donations can be offered in the name of the respondents or even raffles can be promoted.

RS03. *Compose the invitation message.* The invitation message should clearly characterize the researchers involved, the research context and present the recruitment parameters.

R38. In individual messages, introduce persuasive factors that can stimulate the subject participation, such as a compliment and an observation regarding the relevance of each subject' participation.

R39. In on-line surveys, it is highly recommended to send an exclusive token for each participant in order to avoid noise on repeated participations or even the unauthorized distribution of the survey.

6.2.8. Design the Unit of Analysis/Subject Characterization Questions

This activity aims at establishing the additional data that should be gathered out from the source of population to characterize each instance of the unit of analysis/ subject. Typically, the questions regarding the characterization of the survey subjects are included in a specific session of the survey questionnaire. You should perform the following tasks in the presented sequence. Table 14 describes the only recommendation designed for this activity.

CH01. *Identify the unavailable unit of analysis/subjects' attributes.* Identify which unit of analysis/ subject attributes needed are not provided by the source of population or even not updated.

CH02. *Compose the Characterization questions.*

R40. Simple and optional open questions can bring an additional contribution to trace subjects' profile without overloading them with several questions/options, especially when there are few data available and updated in the source of population. However, have in mind that the use of

open questions will restrict quantitative data analysis and the coding process in large scale can be exhaustive.

6.3. Conclusions

This Chapter presented the conceptual framework v3 evolved from the findings obtained in the empirical studies conducted with its previous version (Chapter 5). The conceptual framework activities, tasks and recommendations were significantly evolved although its original scope was preserved. We understand that such version has enough maturity to be evaluated in the field by researchers interested on designing new surveys in different SE topics. Furthermore, taking into account the amount of resources provided by the conceptual framework v3 (including 40 recommendations), we plan to provide a dynamic environment to support its use.

7 Conclusions and Future Work

In this chapter, we present the conclusions, emphasizing the main contributions of this research. Additionally, we present its main limitations and future work addressed to the research questions.

7.1. Final Remarks

Despite the frequent use of survey research in SE, a few discussions have been provided about how to overcome the common lack of samples' representativeness typically caused by different issues observed in planning activities, such as the unclear characterization of the survey target audience and its units, the use of convenience on sampling and the ineffectiveness of the recruitment activities applied. Our investigation allowed us to observe that guidance to overcome such limitation is desirable and can be useful in the field. In this sense, this thesis presented a conceptual framework to support researchers to identify representative samples for surveys in SE.

The third and current version of the conceptual framework was conceived based on findings from individual cases and *in vitro* empirical studies conducted over its previous versions. By adding specific activities, tasks and recommendations to the typical survey planning process, the conceptual framework systematically guides the survey target audience characterization, the identification, the assessment and the selection of available sources of population, the establishment of suitable sampling frames, the establishment of procedures to identify representative samples and the planning of their recruitment.

We are currently inviting partners from different SE research groups to use the conceptual framework in the "battlefield", i.e., to plan surveys addressed to their current researches. Such researches include different SE topics from those already used in our empirical evaluations. After designing their survey plans, the researchers are answering a follow up questionnaire to present their impressions regarding their experience on using of the conceptual framework.

7.2. Contributions of the Presented Research

We observe the following main contributions of the presented research to Empirical Software Engineering research and to SE research in general:

1. A conceptual framework to support researchers to systematically identify representative samples for surveys in SE, composed of:
   - a set of concepts addressed to sampling and recruitment issues;
   - the adaptation of the survey planning process with specific activities and tasks to guide the use of these concepts;
   - a set of 40 recommendations to support the execution of survey planning tasks;
2. A set of experience reports exploring the use of professional social networks as a source of population for investigating different SE topics (DE MELLO and TRAVASSOS, 2013b; DE MELLO, DA SILVA and TRAVASSOS, 2015; DE MELLO, STOLEE and TRAVASSOS, 2015);
3. An investigation on the benefits and drawbacks of using different web-based tools such as professional social networks and crowdsourcing tools to identify representative samples in SE research (DE MELLO and TRAVASSOS, 2012; DE MELLO et al., 2014c; DE MELLO, STOLEE and TRAVASSOS, 2015b);
4. Preliminary findings on how SE survey samples have been established (DE MELLO and TRAVASSOS, 2015) and recruited (DE MELLO and TRAVASSOS, 2016).
5. The protocol of a feasibility study conducted to evaluate the conceptual framework (DE MELLO and TRAVASSOS, 2016);
6. An experience report on the use of focus group sessions to investigate in depth a technology previously evaluated through experiments (DE MELLO and TRAVASSOS, 2016; FRANÇA et al., 2015);
7. The re-execution of a survey on requirement effort influence factors, strengthening evidence on factors that should be taken into account when estimating the effort involved in conducting software requirements activities (DE MELLO and TRAVASSOS, 2013b);
8. The re-execution of a survey on characteristics of agility and agile practices in software processes (DE MELLO, DA SILVA and TRAVASSOS, 2015), allowing us to evolve a body of knowledge to introduce agility in software processes (DE MELLO, DA SILVA and TRAVASSOS, 2014c; DE MELLO and TRAVASSOS, 2016b);
9. The identification of opportunities to improve guidelines for simulation-based studies in SE (FRANÇA and TRAVASSOS, 2015);
10. The replication of an online experiment on Java Code Search, strengthening evidence on how programmers perceive the relevance of different code

snippets provided by different search algorithms on solving Java programming tasks (DE MELLO, STOLEE and TRAVASSOS, 2015).

7.3. Limitations

The conceptual framework presented in this thesis does not intend to provide definitive solutions such as a list of attributes to characterize survey subjects in different research topics. In fact, we initially had planned to support specific SE research topics. However, after investigating the state of practice we observed that basic points from survey research should have been clarified and tailored to SE research in general. Thus, we understood that a less specific technology could be more useful to the community. In addition, the considerable diversity of contexts observed in the field may require different subject characterization to different investigations in the same research topic.

Furthermore, this conceptual framework does not intend to provide a list of sources to be used to search for suitable populations, since such sources could be unstable, typically not designed to support survey research as exemplified in the studies presented in this thesis. Besides, there are survey topics involving audiences eventually available in pretty specific sources.

It is also important to point out that the presented technology does not intend to support the whole survey process, nor even all the survey planning activities. For instance, it does not provide guidance to establish the survey research objective, since we understand it transcends the technology purpose. It also does not provide guidance to design the survey questionnaire, although its consistence with the research objective and target audience and its brevity would influence on the subjects' participation.

Despite the spread of survey research in the field, we highlight that few works are concerned with samples' representativeness and apply more complex sampling designs. Thus, many of the recommendations reported in the conceptual framework are grounded in our own experience to conduct different surveys. Our expectation is that the evolution of the field will allow the conceptual framework to improve with more precise and specific recommendations.

Finally, due to restrictions of time/scope of the presented thesis we did not design/adapt a specific automated environment to support applying the conceptual framework which definitively could influence on the acceptance of the technology in a positive way. One can see that the use of survey tools by researchers is very common, but typically driven to the survey questionnaire design.

7.4. Future Work

In the context of this research, we plan future works, some of them addressed to our initial research questions. Thus, it is important to present them clearly.

How to systematize all the sampling and recruitment activities in order to make them repeatable? As an immediate consequence, we expect to improve the easiness of using the conceptual framework and better exploring its usability by providing a web-based environment to support its instantiation. For instance, such environment will guide the researcher to perform only the tasks and follow the recommendations that are pertinent to the survey characteristics provided. Moreover, the description of the conceptual framework concepts will be more easily visualized when using hypertext resources as proposed by one researcher that used it. In addition, the environment could also allow researchers to reuse previous survey plans to support new investigations and share knowledge regarding the sources of population/search units investigated and successfully applied. One approach under study to develop such environment is providing support to all survey planning and execution activities by integrating the conceptual framework content to resources of questionnaire design and publishing, typically provided by open survey tools, such as LimeSurvey.

How to characterize samples for surveys in SE? / How to stimulate participation in surveys in SE? Another investigation step is concerned with extending the presented structured review to a comprehensive SLR, evolving the investigation on how units of analysis and subjects have been characterized and how subjects have been recruited in surveys from the SE field. Specifically regarding the effect of using persuasive factors on the participation rates, we plan to perform a more controlled investigation by re-executing several times specific surveys following the conceptual framework but applying different recruitment strategies.

Finally, one can see that the presented research was strongly motivated by the observed lack of adequate sampling frames available in the SE field. The presented conceptual framework intends to be an alternative to overcome such limitation by guiding researchers during their research, evaluating and selecting sources of population and supporting the establishment of suitable sampling frames. A second (and ambitious) alternative resides on providing a stable and controlled environment for registering and accessing volunteers able to participate in SE surveys. In this sense, a future work includes designing a collaborative sharing center of volunteers for surveys in SE, initially addressed to registering SE researchers and practitioners.

REFERENCES

ABRANTES, J.; TRAVASSOS, G. (2013) Towards Pertinent Characteristics of Agility and Agile Practices for Software Processes. CLEI Electronic Journal 16(1), article n. 5.

AURUM, A.; PETERSSON, H.; WOHLIN, C. (2002) State of the art: software inspections after 25 years. Software Testing, Verification and Reliability, 12(3), 133-154.

BACKSTROM, C. H.; HURSH-CÉSAR, G. (1981) Survey Research. 2nd Ed. New York, NY: Macmillan Publishing Company, 1981 *apud* KASUNIC, M. (2005) Designing an Effective Survey. TR CMU/SEI-2005-HB-004, Carnegie Mellon University.

BASILI, V. (1993) The experimental paradigm in software engineering. Experimental Software Engineering Issues: Critical Assessment and Future Directions, pp. 1-12. Springer Berlin Heidelberg.

BEGEL, A.; NAGAPPAN, N. (2008) Pair Programming: What's in It for Me? Proceedings of 2nd ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (EASEM), p. 120–128.

BETTENBURG, N.; JUST, S; SCHRÖTER, A.; WEISS, C.; PREMRAJ, R.; ZIMMERMANN, T. (2008) What Makes a Good Bug Report? Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 308-318, 2008.

BIFFL, S.; GRÜNBACHER, P.; HALLING, M. (2006) A family of experiments to investigate the effects of groupware for software inspection. Automated Software Engineering, 13(3), 373-394.

BRAITHWAITE, D.; EMERY, J.; DE LUSIGNAN, S.; SUTTON, S. (2003) Using the Internet to conduct surveys of health professionals: a valid alternative? Family Practice, 20(5), 545-551.

CARVER J. C., HASSLER E., HERNANDES E., KRAFT N. A. (2013). Identifying barriers to the systematic literature review process. In: Proceedings of 7th ACM/IEEE ESEM, p 203–212

CAVALCANTI, Y. C., NETO, P. A. M. S.; MACHADO, I. C.; DE ALMEIDA, E. S., MEIRA, S. R. L. (2013) Towards understanding software change request assignment: A Survey

with Practitioners. Proceedings of 17th International Symposium o Empirical Assessment in Software Engineering (EASE), pp 195–206.

CHEN, J. C.; HUANG, S. J. (2009) An empirical analysis of the impact of software development problem factors on software maintainability. Journal of Systems and Software, 82(6), 981-992.

CIOLKOWSKI, M.; LAITENBERGER, O.; VEGAS, S.; BIFFL, S. (2003) Practical experiences in the design and conduct of surveys in empirical software engineering. In: Conradi, R.; Wang, A. (eds). Empirical Methods and Studies in Software Engineering-Experiences from ESERNET. Springer Berlin Heidelberg, 104-128.

COHEN, L.; MANION, L.; MORRISON, K. (2013) Research methods in education. Routledge.

CONRADI, R.; LI, J.; SLYNGSTAD, O.; KAMPENES, V.; BUNSE, C., MORISIO, M., TORCHIANO, M. (2005) Reflections on conducting an international survey of software engineering. Proceedings of 5th ACM/IEEE International Symposium on Empirical Software Engineering, 214-223, 2005.

DE MELLO, R. M.; PEREIRA, W. M.; TRAVASSOS, G. H. (2010) Activity Diagram Inspection on Requirements Specification. Proceedings of 2010 Brazilian Symposium on Software Engineering (SBES), IEEE. (In Portuguese)

DE MELLO, R. M.; TRAVASSOS, G. H. (2012) Crowdsourcing and Primary Studies on Software Engineering-expectations and challenges. Proceedings of Third Workshop of the Brazilian Institute for Web Science Research, Rio de Janeiro, Brazil. (In Portuguese)

DE MELLO, R. M.; TRAVASSOS, G. H. (2013a) An Ecological Perspective Towards the Evolution of Quantitative Studies in Software Engineering**.** Proceedings of the 17th EASE, p. 216-219.

DE MELLO, R. M.; TRAVASSOS, G. H. (2013b) Would Sociable Software Engineers Observe Better? Proceedings of the 7th ACM/IEEE ESEM, p. 279-282.

DE MELLO, R. M.; DA SILVA, P. C.; TRAVASSOS, G. H. (2014a) Investigating Probabilistic Sampling Approaches for Large-Scale Surveys in Software Engineering. Proceedings of 11th CIBSE/ Workshop on Experimental Software Engineering (ESELAW).

DE MELLO, R. M.; DA SILVA, P. C.; TRAVASSOS, G. H. (2014b) Sampling improvement in software engineering surveys. Proceedings of the 8th ACM/IEEE ESEM, 13-17.

DE MELLO, R. M.; DA SILVA, P. C; TRAVASSOS, G. H. (2014c) Agilidade em Processos de Software: Evidências Sobre Características de Agilidade e Práticas Ágeis. In: XIII Brazilian Symposium on Software Quality, Blumenau, Brazil. (In Portuguese)

DE MELLO, R. M.; DA SILVA, P. C; RUNESON, P.; TRAVASSOS, G. H. (2014a) Towards a Framework to Support Large Scale Sampling in Software Engineering Surveys. Proceedings of the 8th ACM/IEEE ESEM, p. 48:1–48:4.

DE MELLO, R. M., et al. (2014b) Verification of Software Product Line Artefacts: A Checklist to Support Feature Model Inspections. J.UCS 20.5: 720-745.

DE MELLO, R. M.; TRAVASSOS, G. H. (2015a) Characterizing Sampling Frames in Software Engineering Surveys. Proceedings of XVIII Ibero-American Conference on Software Engineering (ESELAW), p. 267-280.

DE MELLO, R. M.; TRAVASSOS, G. H. (2015b) Conceptual Framework to Support Sampling Activities in Software Engineering Surveys. Proceedings of II Latin American School on Software Engineering (ELA-ES), p. 30-41.

DE MELLO, R. M.; DA SILVA, P. C.; TRAVASSOS, G. H. (2015) Investigating Probabilistic Sampling Approaches for Large-Scale Surveys in Software Engineering. Journal of Software Engineering Research and Development, 3(8), 1-26.

DE MELLO, R. M.; STOLEE, K. T.; TRAVASSOS, G. H. (2015) Investigating Samples Representativeness for an Online Experiment in Java Code Search. Proceedings of the 9th ACM/IEEE ESEM, Beijing, 2015.

DE MELLO, R. M.; TRAVASSOS, G. H. (2016) Conceptual Framework for Supporting Sampling and Recruitment in Software Engineering Surveys. Journal of Empirical Software Engineering (submitted).

DE MELLO, R. M.; TRAVASSOS, G. H. (2016b) Beyond Stories, Games and Stand Ups: an Survey on How to Introduce Agility into Software Processes. 24th ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE). (submitted).

DIAS-NETO, A. C.; TRAVASSOS, G. H. (2008) Surveying model based testing approaches characterization attributes. Proceedings of the 2nd ACM/IEEE ESEM, p. 324-326.

DIEBOLD, P; LAMPASONA, C; ZVERLOV, S; VOSS, S. (2014) Practitioners' and researchers' expectations on design space exploration for multicore systems in the automotive and avionics domains: a survey. Proceedings of 18th EASE, 10pp.

DIEBOLD, P.; VETRÒ, A. (2014) Bridging the Gap: SE Technology Transfer into Practice–Study Design and Preliminary Results. Proceedings of the 8th ACM/IEEE ESEM, article n.15.

DYBÅ, T.; SJØBERG, D. I.; CRUZES, D. S. (2012) What works for whom, where, when, and why?: on the role of context in empirical software engineering. Proceedings of the 6th ACM-IEEE ESEM, p. 19-28.

eSEE (2015) Experimental Software Engineering Environment: Systematic Reviews Studies. http://ese.cos.ufrj.br/wikiese/index.php/Systematic_Reviews_Studies. Accessed 23 January 2016.

ESE (2015) Experimental Software Engineering Group. http://ese.cos.ufrj.br/ese/

FÆGRI, T.; DYBÅ, T.; DINGSØYR, T. (2010) Introducing knowledge redundancy practice in software development: Experiences with job rotation in support work. Information and Software Technology, 52(10), 1118-1132.

FERNÁNDEZ-SÁEZ, A. M.; GENERO, M.; CAIVANO, D.; CHAUDRON, M. R. (2016) Does the level of detail of UML diagrams affect the maintainability of source code?: a family of experiments. Empirical Software Engineering, 21.1 (2016) 1-48.

FERNANDÉZ-SÁEZ, A. M.; CHAUDRON, M.; GENERO, M.; RAMOS, I. (2013) Are forward designed or reverse-engineered UML diagrams more helpful for code maintenance?: a controlled experiment. Proceedings of the 17th ACM EASE, p. 60-71.

FRANÇA, A.; DA SILVA F. An Empirical Study on Software Engineers Motivational Factors. In: Proceedings of the 3rd ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 405–409, 2009.

FRANÇA, A.; DA SILVA F.; MARIZ, L. (2010) An Empirical Study on the Relationship between the Use of Agile Practices and the Success of Scrum Projects. Proceedings of ACM/IEEE 4th ESEM, 37:1–37:4, 2010.

FRANÇA, B. B. N.; TRAVASSOS, G. H. (2015) Experimentation with dynamic simulation models in software engineering: planning and reporting guidelines. Empirical Software Engineering, 1-44.

FRANÇA B. B. N.; RIBEIRO, T. V.; MEDEIROS, P. S.; DE MELLO, R. M.; TRAVASSOS, G. H. (2015) Using Focus Group to Support Software Engineering Research. Journal of Software Engineering Research and Development (submitted)

GLYNN, E.; FITZGERALD, B.; EXTON, C. (2005) Commercial adoption of open source software: an empirical study. Proceedings of 5th ACM/IEEE ISESE.

GOETZ, E.; TYLER, T.; COOK, F. (1984) Promised Incentives in Media Research: A Look at Data Quality, Sample Representativeness, and Response Rate. Journal of Marketing Research. 21(2), 148-154.

GROVES, R. M.; FOWLER JR, F. J.; COUPER, M. P.; LEPKOWSKI, J. M.; SINGER, E.; TOURANGEAU, R. (2011). Survey methodology (Vol. 561). John Wiley & Sons.

GUO, Y., SEAMAN, C. (2008) A survey of software project managers on software process change. Proceedings of the 2nd ACM/IEEE ESEM. ACM, p. 263-269.

GUZMÁN L.; LAMPASONA C.; SEAMAN C.; ROMBACH, D. (2014) Survey on research synthesis in Software Engineering. Proceedings of 18th EASE, article n. 2, 10pp.

HAIR, J.; BLACK, W.; BABIN, B.; ANDERSON, R. (2009) Multivariate Data Analysis, 7th ed. Prentice Hall.

HANENBERG, S.; KLEINSCHMAGER, S.; ROBBES, R.; TANTER, É.; STEFIK, A. (2014). An empirical study on the impact of static typing on software maintainability. Empirical Software Engineering, 19(5), 1335-1382.

HOGGANVIK, I.; STOLEN, K. (2005) Risk analysis terminology for IT-systems: does it match intuition? Proceedings of the 5th ACM/IEEE International Symposium on Empirical Software Engineering (ISESE).

HOPKINS, K. (1982) The Unit of Analysis: Group Means Versus Individual Observations. American Education Research Journal, 19: 5–18.

HÖST, M.; ORUČEVIĆ-ALAGIĆ, A.; RUNESON, P. (2011) Usage of open source in commercial software product development–findings from a focus group meeting. Product-Focused Software Process Improvement, p. 143-155, Springer Berlin Heidelberg.

HUMAYUN, M.; GANG, C.; MASSOD, I. (2013) An empirical study on investigating the role of KMS in promoting trust within GSD teams. Proceedings of the 17th ACM EASE, p. 207-211.

JI, J.; LI, J.; CONRADI, R., LIU, C., MA, J., CHEN, W. (2008) Some Lessons Learned in Conducting Software Engineering Surveys in China. Proceedings of the 3nd ACM/IEEE ESEM, p. 168–177.

JOHNS, G. (1991) Substantive and methodological constraints on behavior and attitudes in organizational research. Organizational Behavior and Human Decision Processes, 49: 80-104.

JOORABCHI, M; MESBA, A.; KRUCHTEN, P. (2012) Real challenges in mobile app development. Proceedings of the 7th ACM/IEEE ESEM, 15-24.

JURISTO, N.; MORENO, A.; VEGAS, S. (2004) Reviewing 25 years of testing technique experiments. Empirical Software Engineering, 9(1-2), 7-44.

KANIJ, T.; MERKEL, R.; GRUNDY, J. A. (2011) Preliminary Study on Factors Affecting Software Testing Team Performance. Proceedings of 5th ACM/IEEE ESEM, p. 359–362.

KARLSSON, J.; WOHLIN, C; REGNELL, B. (1998) An evaluation of methods for prioritizing software requirements. Information and Software Technology 39.14 (1998): 939-947.

KASURINEN, J.; TAIPALE, O.; SMOLANDER, K. (2010) Test Case Selection and Prioritization: Risk-based or Design-based? Proceedings of 4th ACM/IEEE ESEM, p. 10:1–10:10.

KASUNIC, M. (2005) Designing an Effective Survey. TR CMU/SEI-2005-HB-004, Carnegie Mellon University. http://www.sei.cmu.edu/reports/05hb004.pdf. Accessed 20 Nov 2014.

KITCHENHAM, B. PFLEEGER, S. (2008) Personal Opinion Surveys. In: Shull, F.; Singer, J.; Sjøberg, D. (eds.) Guide to Advanced Empirical Software Engineering. 63–92. Springer London.

KRUSKAL, F.; MOSTELLER, F. (1979) Representative Sampling III: The Current Statistical Literature. International Statistical Review 47(3), 245-265..

KUHN, D.; WALLACE, D.; GALLO, J. (2004) Software fault interactions and implications for software testing. IEEE Transactions on Software Engineering, 30(6), 418-421.

KUPPER, L.; HAFNER, K. (1989)How appropriate are popular sample size formulas? The American Statistician, 43(2), 101-105.

KUSUMO, D.; STAPLES, M.; LIMING, Z.; JEFFERY, R. (2012) Analyzing differences in risk perceptions between developers and acquirers in OTS-based custom software projects using stakeholder analysis. Proceedings of the 6th ACM/IEEE ESEM, p. 69-78.

LEVIN, J.; FOX, J.; FORDE, D. (2013) Elementary Statistics in Social Research. 11th Edition, Pearson Education.

LINÅKER J.; SULAMAN S. M.; DE MELLO R. M.; HÖST, M. (2015) Guidelines for conducting surveys in Software Engineering. TR 5366801, Lund University Publications.

MOE, N. B.; ŠMITE, D; ŠĀBLIS, A; BÖRJESSON. A. L.; ANDRÉASSON, P. (2014) Networking in a large-scale distributed agile project. Proceedings of 8th ACM/IEEE ESEM, article n. 12, 9pp.

MONASOR, M. J.; NOLL, J.; VIZCAÍNO, A.; PIATTINI, M.; BEECHAM, S. (2014) Walk before you run: using heuristic evaluation to assess a training tool prototype. Proceedings of 18th EASE, article n. 41, 10pp

MURPHY, B.; BIRD, C.; ZIMMERMANN, T.; WILLIAMS, L.; NAGAPPAN, N.; BEGEL, A. (2013) Have Agile Techniques been the Silver Bullet for Software Development at Microsoft? Proceedings of 7th ACM/IEEE ESEM, p. 75–84.

NGUYEN-HOAN, L.; FLINT, S.; SANKARANARAYANA, R. (2010) A Survey of Scientific Software Development. In: Proceedings of 4th ACM/IEEE ACM/IEEE ESEM, p. 12:1–12:10.

NUGROHO, A.; CHAUDRON, M. (2007) A Survey of the Practice of Design – Code Correspondence amongst Professional Software Engineers. Proceedings of 2nd ACM/IEEE ESEM, p. 467–469.

NUGROHO, A.; CHAUDRON, M. A. (2008) Survey into the Rigor of UML Use and Its Perceived Impact on Quality and Productivity. Proceedings of 2nd ACM/IEEE ESEM, p. 90–99.

NUNNENMACHER, S.; JUNG, J.; CHEHRAZI, G.; KLAUS, A.; LAMPASONA, C.; WEBEL, C.; CIOLKOWSKI, M. A. (2011) Preliminary Survey on Subjective Measurements and Personal Insights into Factors of Perceived Future Project Success. Proceedings of 5th ACM/IEEE ESEM, 396–399, 2011.

ODZALY, E.; GREER, D.; SAGE, P. (2009) Software risk management barriers: An empirical study. Proceedings of 3rd ACM/IEEE ACM/IEEE ESEM, p. 418–421.

PINSONNEAULT, A.; KRAEMER, K. (1993) Survey Research Methodology in Management Information Systems: An Assessment. Journal of Management Information Systems 10, (2), 75-105.

PFAHL, D.; YIN, H.; MÄNTYLÄ, M. V.; MÜNCH, J. (2014) How is Exploratory Testing Used? A State-of-the-practice Survey. Proceedings 8th ACM/IEEE ESEM, article n. 5, 10pp.

PFLEEGER, S.; KITCHENHAM, B. (2001) Principles of survey research part 1: Turning Lemons into Lemonade. ACM SIGSOFT Software Engineering Notes 26(6), 16-18.

PFLEEGER, S.; KITCHENHAM, B. (2002) Principles of survey research: part 5: populations and samples. ACM SIGSOFT Software Engineering Notes, 27(5), 17-20.

PUNTER ET AL., Conducting On-line Surveys in Software Engineering. Proceedings of the ACM/IEEE 2003 International Symposium on Empirical Software Engineering, 80-88, 2003.

RAFI, D. M.; MOSES, K. R. K.; PETERSEN, K.; MÄNTYLÄ, M. V. (2012). Benefits and limitations of automated software testing: Systematic literature review and practitioner survey. In Proceedings of the 7th International Workshop on Automation of Software Test, pp. 36-42, IEEE Press.

RICCA, F., et al. (2009) Using acceptance tests as a support for clarifying requirements: A series of experiments. Information and Software Technology 51.2 (2009): 270-283.

RODRÍGUEZ, P. et al (2013) Building lean thinking in a telecom software development organization: strengths and challenges. Proceedings of the 2013 International Conference on Software and System Process, ACM, pp 98-107.

RODRÍGUEZ, P.; MARKKULA, J.; OIVO, M.; TURULA, K. (2012) Survey on agile and lean usage in Finnish software industry. Proceedings of the 6th ACM/IEEE ESEM, p. 139-14.

RUNESON, P.; HÖST, M. (2008) Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering 14(2), 131-164, 2008.

SANTA IZABEL S. L., TRAVASSOS G. H. (2011) Características de técnicas de teste de software para uso em projetos Web. Proceedings of 14th Ibero-American Conference on Software Engineering (CibSE), 14pp.

SANTOS, R. E. S.; DA SILVA F. Q. B. (2013) Motivation to perform systematic reviews and their impact on Software Engineering practice. Proceedings of 7th ACM/IEEE ESEM, p. 292–295.

SÄRNDAL, C.; SWENSSON, B.; WRETMAN, J. (1992) Model Assisted Survey Sampling. Springer.

SENAPATHI, M.; SRINIVASAN, A. (2014) An Empirical Investigation of the Factors Affecting Agile Usage. Proceedings of 18th EASE, 10pp.

SHNEIDERMAN, B. (2013) High Impact Research: Blending Basic and Applied Methods. Proceedings of 7th ACM/IEEE ESEM (Keynote Speech), 3pp..

SJØBERG, D; HANNAY, J.; HANSEN, O; KAMPENES, V.; KARAHASANOVIC, A.; LIBORG, N.; REKDAL, A. (2005) A survey of controlled experiments in software engineering. IEEE Transactions on Software Engineering 31(9): 733-753.

SMITH, E.;LOFTIN, R.; MURPHY-HILL, E.; BIRD, C.; ZIMMERMANN, T. (2013) Improving developer participation rates in surveys. In: Proceedings of 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE), IEEE, pp 89-92.

SPÍNOLA R. O.; Travassos G. H. (2012) Towards a framework to characterize ubiquitous software projects. Information and Software Technology 54(7):759-785.

STETTINA, C. J., AND HEIJSTEK, W. (2011) Five agile factors: Helping self-management to self-reflect. Systems, Software and Service Process Improvement. Springer Berlin Heidelberg, p. 84-96.

STOLEE, K.; ELBAUM, S.; DWYER, M. (2015) Code Search with Input/Output Queries: Generalizing, Ranking and Assessment. The Journal of Systems and Software.

STAVRU, S. (2014) A critical examination of recent industrial surveys on agile method usage. Journal of Systems and Software 94: 87-97.

SLYNGSTAD, O.; GUPTA, A.; CONRADI, R.; MOHAGHEGHI, P.; RØNNEBERG, H., LANDRE, E. (2006) An Empirical Study of Developers Views on Software Reuse in Statoil ASA. Proceedings of 6th ACM/IEEE ISESE, p. 242–251.

SOUTHEKAL, P.; LEVIN, G. (2011) Formulation and Empirical Validation of a GQM Based Measurement Framework. Proceedings of 5th ACM/IEEE ESEM, p. 404–413.

TAIPALE, O., SMOLANDER, K. (2006) Improving Software Testing by Observing Practice. In: Proc. 6th ACM/IEEE ISESE, p. 262–271.

TAIPALE, O., KARHU, K., SMOLANDER, K. (2007) Observing Software Testing Practice from the View-point of Organizations and Knowledge Management. Proceedings of 1st ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 21–30.

THOMPSON, S. (2012) Sampling. John Wiley & Sons, 3rd edition.

TOFAN, D.; GALSTER, M.; AVGERIOU, P.; WEYNS, D. (2011) Software engineering researchers' attitudes on case studies and experiments: An exploratory survey. Proceedings of 15th EASE, p. 91–95.

TOMASSETI, F.; TORCHIANO, M.; TISO, A.; RICCA, F.; REGGIO, G. (2012) Maturity of software model-ling and model driven engineering: A survey in the Italian industry. In: Proc. 16th EASE, p. 91–100.

TORCHIANO, M.; TOMASSETTI, F.; RICCA, F.; TISO, A.; REGGIO, G. (2011) Preliminary Findings from a Survey on the MD State of the Practice. Proc. 5th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, p. 372–375.

TRAVASSOS, G.; SHULL, F.; FREDERICKS, M.; BASILI, V. R. (1999) Detecting defects in object-oriented designs: using reading techniques to increase software quality. ACM Sigplan Notices, 34(10), 47-56. ACM.

VAZ, V. (2013) Software Requirements Effort Estimation. Master Dissertation. PESC/COPPE- UFRJ.

WOHLIN, C. ET AL. (2012) Experimentation in Software Engineering. Springer.

WNUK, K.; GORSCHEK, T.; ZAHDA, S. (2013) Obsolete software requirements. Information and Software Technology, 55(6), 921-940.

Appendix A

Feasibility Study: Experimental Task (In Portuguese)

Tarefa- Planning a Survey on Requirements Effort Influence Factors

Estudo sobre o planejamento de surveys em Engenharia de Software

Nome do Participante: _____

Orientações

Este documento apresenta oito itens que devem ser contemplados no plano de um survey sobre fatores de influência do esforço nas atividades de requisitos. Dentre estes itens, apenas o objetivo de pesquisa (item 1) e o questionário (item 8) já estão preenchidos. A tarefa consiste, portanto, em que você preencha os demais itens (2,3,4,5,6 e 7), relacionados ao processo de amostragem e ao recrutamento de participantes.

(Para realizar sua tarefa, pedimos que você utilize o framework encaminhando na mesma mensagem em que você recebeu esta tarefa (framework_v1.pdf). Este framework é composto de conceitos, atividades e recomendações para auxiliar atividades de amostragem e de recrutamento em surveys da Engenharia de Software)[19]

Você pode responder aos itens em português. Não existe uma única resposta "certa" para os itens. Utilize seu conhecimento sobre o tema do survey e as informações contidas no objetivo de pesquisa e no questionário para refletir sobre suas respostas.

Você pode utilizar como apoio o material da aula de surveys e o documento de guidelines encaminhado na mesma mensagem em que recebeu esta tarefa (survey_guidelines.pdf).

Lembre-se que o prazo para entrega deste documento preenchido (via e-mail) é ***04/08/2015*** *(terça-feira)!*

1. Objetivo de Pesquisa

It aims at complementing an investigation concerned with context variables that may affect the effort involved in requirements activities. According to our perspective, the total requirements effort estimation shall be calculated by adding efforts spent on dealing with requirements activities (called effort components). Besides, as usually happen in software projects, each effort component can be influenced by different factors (called influence factors). So, understanding effort components, influence factors and their relationships can increase the likelihood of better requirements effort estimation.

2. Audiência-alvo (*Target audience*)

[19] Paragraph included in the document only to the subjects in the FWK group.

3. Unidade de Análise *(Unit of analysis)*

4. Caracterização do participante (*Participants' attributes*)

5. Quadro de amostragem (*Sampling frame*)

6. Design de amostragem (*Sampling design*) a ser aplicado

7. Descrição detalhada do processo de recrutamento (*Recruitment proccess*)

8. Questionário

1) According to your experience, the total requirements project effort can be calculated by adding the following effort components: *(Check any that apply)*

( ) Effort spent coordinating the technical team and its interaction with stakeholders
( ) Effort spent on changing requirements
( ) Effort spent on communication gaps
( ) Effort spent on control of team performance
( ) Effort spent on knowledge transfer
( ) Effort spent on specifying the requirements

Can you remember any other effort component that can be considered to calculate the total effort? *(Indicate here if you think that that are other effort components involved in requirements activities that were not represented in the previous components list)*

1-
2-
3-
4-
5-

2) According to your experience, indicate which factors can somehow influence the estimation of requirements effort. Please, do not consider the requirements team as stakeholder. (Check any that apply)

( ) Accessibility of information about the problem domain and operative environment
( ) Availability of a defined process (methods, techniques, tools, activities, roles and artifacts) for requirements engineering
( ) Availability of process and configuration management tool for controlling requirements changes
( ) Percentage of effort that must be allocated to quality assurance in relation to the specification effort

( ) Reuse of existing artifacts (templates, requirements of other projects, etc.)
( ) Average team productivity
( ) Clearness of roles defined for team members
( ) Degree of distribution of the technical team
( ) Familiarity of the team with the technologies, tools and platforms to be used
( ) Number of resources in requirements team
( ) Personnel turnover
( ) Quality of communications between team members
( ) Team morale
( ) Team skill/experience
( ) Team understanding about the problem domain
( ) Cultural distance between the client and the technical team (language, punctuality, formality, etc.)
( ) Geographical distance between the client and the technical team
( ) Intensity of communication between requirements team and stakeholders
( ) Quality of interaction between technical team and stakeholders
( ) Existence of divergences or conflicts of interest between stakeholders
( ) Number of stakeholders involved in requirements elicitation
( ) Percentage of key stakeholders involved in requirements elicitation
( ) Reliability of stakeholders
( ) Stakeholder's animosity with the project
( ) Stakeholder's demand with respect to the formalism and level of detail of the artifacts to be generated
( ) Stakeholder's understanding of the techniques and technologies to be used (ability to understand the specifications, diagrams, etc.)
( ) Degree of understanding of the product vision
( ) Natural complexity of problem domain
( ) Stability of business environment
( ) Conceptual compliance between software requirements and the real world
( ) Existence of design patterns to be met
( ) Level of completeness (detailing / formalism) of requirements specification
( ) Novelty of application
( ) Solution size and complexity
( ) Expected project duration
( ) Feasibility of concluding the project in time and on budget
( ) Loose schedule (slack time available)
( ) Time pressure and risk of schedule overrun
( ) Percentage of errors in requirements specification
( ) Risk of change in requirements

Can you remember any other factor that can be considered to calculate the total effort? *(Indicate here if you think that that are other effort components involved in requirements activities that were not represented in the previous components list)*

1-
2-
3-
4-
5-

3) Considering the effort components you have indicated, which one you suggest to be the biggest part of the overall requirements effort?

4) Establish if the effort component that you selected in question 3 perform a *positive* or a *negative* relationship with each influence factor that you indicated in question 2.

We would like to thank you for your collaboration!!! Results of this study will be used only for our research regarding requirements effort influence factors. As soon as we have completed our technical report we will notify all participants. If you would like more information about this research, please email us.

Appendix B

Conceptual Framework v3: Concepts

1. Target audience

A survey *target audience* characterizes who can best provide the information needed in order to achieve the research objective.

2. Unit of analysis and Study subject

In opinion surveys, data is always collected from *subjects* necessarily represented by the each respondent, i.e., an *individual*. However, the survey target audience may demand a different level of analysis, i.e., the *unit of analysis*. In the context of SE surveys, other possible types of unit of analysis include: *organizational unit, organization and project team*. Each instance of unit of analysis/study subject is characterized by a set of *attributes,* including *control attributes*. *Control attributes* followed by predefined values can be used for representing the restrictions previously established by the target audience. For instance, if the target audience restricts the survey to be applied to *Brazilian individuals working as project managers*, "*country= Brazil*" and "*role= project manager*" are control attributes.

3. Population

Survey *population* consists on a set of accessible objects addressed to a specific target audience.

4. Sampling frame

In statistics, a *sampling frame* is the list composed by the objects from the survey population. Thus, the sampling frame is the source from which a *sample*, i.e. a subset of objects from the population can be retrieved. 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. Some appropriate investigations could not be carried out due to the lack of a suitable sampling frame, while other investigations remain inconclusive results due to incomplete sampling frames.

5. Source of Population

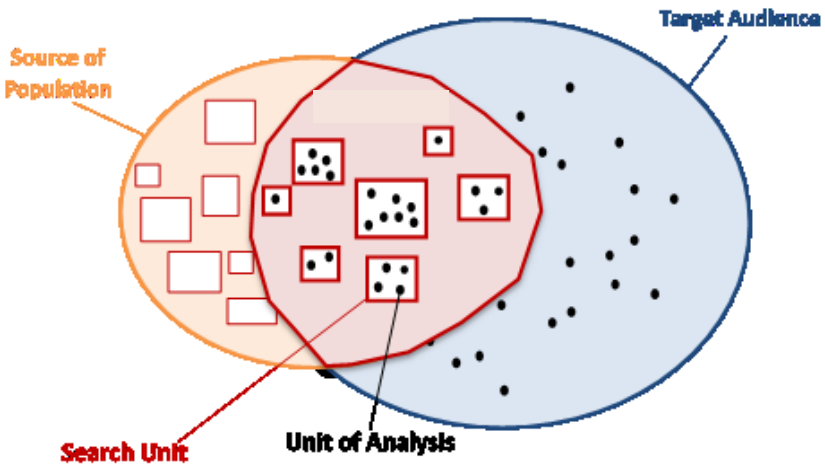
A *source of population* consists on a database (automated or not) from which an adequate population for specific target audience can be systematically retrieved. This, if a source of population can be considered *valid* in a specific research context, it can be concluded that also *valid sampling frames* can be established from it for the same research context.

4.1 Search unit

The *search unit* characterizes the entity from which one or more objects can be retrieved from a specific *source of population*. Indeed, in an ideal scenario, it is expected that both *unit of analysis* and *search unit* are represented by the same entity. However, SE literature presents some examples in which these elements are different, such as the following:

- Conradi et al. (2005) aims at analyzing *project teams*, but accessed them keeping in touch with *organizations* from three distinct countries.
- Dias Neto et Travassos (2008); Carver et al. (2013); Santos et al. (2013); Abrantes and Travassos (2013); Gúzman et al. (2014) opted to survey the authors of the *papers* retrieved from the results of specifics SLRs conducted for each research context.
- De Mello et al. (2014) used *groups of interest* from a professional social network to sample individuals, since the tool significantly restricted the direct access to individuals out from their groups.

Following figure associates the concepts of *source of population* and *search unit* with the concepts of *target audience*, and *unit of analysis*.



4.2 Criteria for Assessing Candidates to Source of Population

To be considered *valid*, a source of population should satisfy, at least, the following *essential requirements* (ER):

- ***ER1.*** A source of population should not intentionally represent a segregated subset from the target audience, i.e., for a target population audience "X", it is not adequate to search for units from a source intentionally designed to compose a specific subset of "X".
- ***ER2.*** A source of population should not present any bias on including on its database preferentially only subsets from the target audience. Unequal criteria for including search units mean unequal sampling opportunities.
- ***ER3.*** All source of population' search units, objects and subjects must be identified by a logical or numerical id.
- ***ER4.*** All source of population' search units, objects and subjects must be accessible. If there are hidden elements, it is not possible to contextualize the population.

There are also nine *desirable requirements* (DR), three concerned with the samples' *accuracy* (ADR), two concerned with *clearness* (CDR) and four regarding sample's *completeness* (CoDR).

- ***ADR1.*** It is possible to retrieve each search unit from the source of population in a logical and systematic way.
- ***ADR2.*** There is no unit of analysis outside the target audience concerned with the source of population.
- ***ADR3.*** There is a one-to-one correspondence between each search unit and each unit of analysis.
- ***CDR1.*** All search units appear once in the source of population.
- ***CDR2.*** All units of analysis appear once in the source of population.
- ***CoDR1.*** All information needed from each search result is up-to-date.
- ***CoDR2.***All information needed from each unit of analysis/subject is accessible and up-to-date.
- ***CoDR3.***All units of analysis from the target audience can be found in the source of population.

- ***CoDR4.*** Each search result provides relevant and sufficient information for performing alternative probabilistic sampling designs, such as stratified and cluster sampling.

4.3 Sources of Population types

Depending on the unit of analysis and the research context, many types of sources can be used, such as:

- *SE Conferences: individuals* assisting to relevant SE conferences can be eventually considered a small but representative set for many surveys contexts, especially when researchers are the target audience (Torchiano et al., 2011; Monasor et al., 2014);
- *Discussion Groups:* an active and thematic SE discussion group can be considered as good source for sampling *individuals* (Nugroho and Chaudron, 2007);
- *Projects repositories:* retrieving a representative dataset from *software projects*, including data from the project team, is a challenge. Typically, data from several projects can be retrieved in the context of open source projects (Bettenbourg et al., 2015).
- *Digital Libraries:* when the target audience are restricted to researchers, Digital libraries such as SCOPUS and IEEE can be used for retrieving relevant authors of *papers* in the survey context (Dias Neto et al., 2008; Carver et al., 2013; Santos and Da Silva, 2013; Abrantes and Travassos, 2013; Gúzman et al., 2014);
- *Catalogues:* searching for National or International catalogues provided by institutes (Rodríguez et al., 2012), government or even yellow pages (Conradi et al., 2005) may be considered for retrieving representative sets of *organizations* or *individuals*;
- *Professional Social Networks:* it has demonstrated to be a promising technology for supporting large scale sampling of *individuals*, that can be directly accessed or through *groups of interest* (de Mello et al., 2015; Torchiano, 2011). However, limitations on accessing searching units must be taken into account;

This list does not intend to be exhaustive and other types of sources can be taken into account. De Mello et al. (2014-3) applied the essential requirements and the desirable requirements in nine distinct sources of sampling available in the Web, including *crowdsourcing tools*, *professional social networks* and *freelancing tools*. While it was observed

that the analyzed freelancing tools and professional social networks could be used as source of population, it was concluded that the analyzed crowdsourcing tools do not support to essential requirements.

5 Population Search Plan

A *population search plan* describes how instances from the survey search unit will be systematically retrieved from the selected source of population and evaluated in order to compose the survey population (de Mello et al., 2014-3).

5.1 Search String

A *search string* is composed by a set of *search expressions* connected through logical operators that can be applied to a source of population in order to retrieve adequate search units. As in the case of systematic literature reviews (SLRs), we argue that search expressions can be applied to avoid bias on filtering the relevant elements from the point of view of the *research objective*. Search strings must be avoided when there are no relevant units of analysis outside the target audience concerned with the source of population (ADR2). This could happen, e.g., when the source of population is composed of the list of employees from a SE organization, and the set of employees from this organization composes the target audience.

5.2 Search Algorithm

The *search algorithm* describes each step, automated or not, that should be followed in order to filter the search units in a source of population, including how to apply the planned search string. A search algorithm can vary significantly on complexity, depending on the resources available in the source of population. In addition, any previously known restrictions for accessing the search units should be described (and how to deal with them).

5.3 Inclusion/Exclusion Criteria

Another concept borrowed from SLRs, the *inclusion/exclusion criteria* describe a set of restrictions that should be applied in order to include/exclude search units retrieved from the search plan execution. Exclusion criteria can be especially helpful when the source of population is significantly generic and the use of search string are limited, such as in the case of the professional social networks (de Mello et al., 2015) and

yellow pages (Conradi et al., 2005). As in the case of search strings, if the requirement ADR2 is satisfied by a source of population, exclusion criteria should be avoided.

6 Sampling Strategy

A *sampling strategy* characterizes how sampling will be performed (*sampling design*) and how many objects should be selected (*sample size*).

6.1 Sampling design

Describes the criteria for extracting samples from the sampling frame, i.e., which objects will be selected to support the survey execution. In general, probabilistic (randomly) sampling approaches are recommended: *simple random sampling*, *clustered sampling*, *stratified sampling* and *systematic sampling*. Eventually, all individuals in the sampling frame can be included in the sample (census).

6.1.1 Non-Probabilistic Sampling

Non-probabilistic sampling addresses to all sampling designs in which randomness cannot be observed on selecting the samples, i.e., the samples from the survey population do not have the same probability to be chosen (Thompson, 2012). As main consequence, the extent in which the observed results can be generalized to all survey population is limited. 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 own personal connections (de Mello and Travassos, 2015-2).
- *Quota sampling:* The sampling frame is composed by mutually exclusive subsets in which quotas of sample size are arbitrary established to each subset. For instance, a survey design could establish that 20 companies will be surveyed limiting to sample 10 employees from each company.
- *Judgment sampling:* It aims to reduce the bias from the accidental sampling providing some criteria for selecting each unit, such as the use of experts' opinion. For instance, França and Travassos (2015) analyzed the list of ISERN

members and the directory of Brazilian researchers (CNPq) in order to find individuals able to participate in their survey.

- *Snowballing sampling:* It extends accidental sampling, typically recruiting an initial set of subjects (seeds, first wave) that will be able to indicate more subjects (second wave) to participate in the survey. It is important to distinguish snowballing from the common practice observed in SE surveys in which any subjects is able to recruit other subjects, out of researchers' control. In such cases, even a sampling frame was not established.

6.2 Probabilistic Sampling

In probabilistic sampling designs, all objects available in the study population have the same probability to be selected. As a consequence, it will be feasible to calculate the *confidence level* of the observed results (in which extent the results are reliable) and its *confidence interval* (in which level the results can be extended to all population).

6.3 Simple Random Sampling

In simple random sampling, each member of the population under study has an equal chance of being selected and the probability of a object of the population being selected is unaffected by the selection of other objects from the population, i.e. each selection is entirely independent of the next. Thus, performing SRS indicates that all its objects can be considered *homogeneous* from the point of the view of the study scope. For instance, Rodríguez et al (2005) randomly selected a subset of members from FIPA (The Finnish Information Processing Association) to participate in their survey.

6.4 Systematic Sampling

Systematic sampling consists on a SRS in which a population composed by N objects have a sample of n objects selected following a sequence initialized by a randomly selected unit i. Then, the next objects 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 objects will be included in this sample: 3, 7, 11, 15, 19, 23, 27, 31, 35 and 39.

6.5 Clustered Sampling

In this sampling design, homogeneous clusters composed by distinct but similar objects can be identified in a single population. Thus, the similarity observed allow researchers sampling only from subset from these clusters. As a consequence, a certain loss of confidence in the results is expected but, at the same time, significant efforts on recruiting and data collecting could be avoided. Thus, clustered sampling is commonly applied in large scale surveys in which researchers must be *in loco*, (in person) for gathering data. Due to the already discussed issues regarding sampling in SE surveys, the identification of clusters in a population can be considered a great challenge.

6.6 Stratified Sampling

Stratified sampling is considered the best sampling design to support large scale studies, distributing all the population objects into distinct subpopulations (*strata*) naturally available. Then, for each stratum, SRS must be performed, allowing observing more specific and reliable results than in a single population. It is important to highlight that no objects from a sampling frame can be let out from a stratum and all strata must be mutually exclusive, i.e. a single unit cannot be found in more than one strata. However, our investigations did not identified previous surveys in SE using stratified sampling.

6.7 Sample Size

When using a probabilistic sampling design, the study *sample size* should be calculated in function of the survey *population size*, the aimed *confidence interval* and *confidence level*. *Confidence level* is an index of how sure we can be that survey responses will lie in a given variation range, i.e., a specific *confidence interval* (COHEN, MANION and MORRISON,

2013). It is commonly established a confidence level of 95% or 99% while the following formula can be used to calculate the sample size (Kupper *et al.*, 1989):

$$SS = \frac{Z^2 \times p \times (1-p)}{c^2}$$, where:

- *SS= sample size*
- *Z= Z-value, established through a specific table (Z=2.58 for 99% of confidence level, Z=1.96 for 95% of confidence level).*
- *p= percentage picking a choice, expressed as decimal (0.5 used as default for calculating sample size, since it represents the worst case).*
- *c= desired confidence Interval, expressed in decimal points (Ex.: 0.04).*

For the calculation of a sample size based on a *finite* population with a *pop* size, the following correction formula should be applied over *SS*:

$$SS_f = \frac{SS}{1 + \frac{SS-1}{pop}}$$

For instance, considering a population composed by 10.000 individuals in which is acceptable that the observed results could vary in ±10 points (confidence interval) for a confidence level of 95%, a sample composed at least by 95 individuals is needed. However, in this example, sampling only 95 individuals is only recommended if you can be sure that all subjects will effectively participate in the study. In surveys, participation if frequently voluntary and higher sample sizes should be established to mitigate the effect of the *participation rate*. For instance, if experience on previous study executions indicates that only 20% of the subjects tend to effective participate in the exemplified study, it can be considered a good practice to recruit a sample size five times higher than the calculate sample size (475).

7 Recruitment Strategy

The recruitment strategy characterizes how the individuals from the survey sample will be recruited. It includes the *invitation message* and the following factors that can influence subjects' participation (Smith et al., 2013).

- *Execution esteemed time:* the mean time esteemed to each subject fill out the survey. It can be calculated based in the results of pilot executions;

- *Invitation method:* characterizes how the invitation message will be sent. In the case of invitations supported through Web, common approaches are *sending individual and personalized e-mails; individual and generic e-mails; sending a single generic e-mail; sending a single generic message to a list or a group*.
- *Period Available:* characterize how for many time (typically in days) the survey will be available for the subjects.
- *Reminding method:* any intention of reminding individuals regarding answering the survey should be described. For instance, it can be planned to remind the subjects once, re-sending the original message after one week of the survey execution.
- *Reward method:* it must be described if the subjects can be stimulated through any kind of reward and in which case it will be offered (e.g. if only the subject complete the survey questionnaire). In this context, rewards may but is not limited to include *payments*, *raffles*, *gifts* and *donations for NGOs*.