Recruiting credible participants for field studies in software engineering research

Austen Rainer[a], Claes Wohlin[b]

[a]*Queen's University Belfast, 18 Malone Road, Computer Science Building, BT9 5BN, Belfast, Northern Ireland, UK*
[b]*Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden*

Abstract

Context: Software practitioners are a primary provider of information for field studies in software engineering. Research typically recruits practitioners through some kind of sampling. But sampling may not in itself recruit the "right" participants.
Objectives: To assess existing guidance on participant recruitment, and to propose and illustrate a framework for recruiting professional practitioners as credible participants in field studies of software engineering.
Method: We review existing guidelines, checklists and other advisory sources on recruiting participants for field studies. We develop a framework, partly based on our prior research and on the research of others. We search for and select three exemplar studies (a case study, an interview study and a survey study) and use those to illustrate the framework.
Results: Whilst existing guidance recognises the importance of recruiting participants, there is limited guidance on how to recruit the "right" participants. The framework suggests the conceptualisation of participants as "research instruments" or, alternatively, as a sampling frame for items of interest. The exemplars suggest that at least some members of the research community are aware of the need to carefully recruit the "right" participants.
Conclusions: The framework is intended to encourage researchers to *think differently* about the involvement of practitioners in field studies of software engineering. Also, the framework identifies a number of characteristics not explicitly addressed by existing guidelines.

Keywords:
Credibility, Validity, Reliability, Data collection, Sampling, Subjects, Participants, Recruitment

1. Introduction

In this article, we propose and illustrate a framework for recruiting credible participants for field studies of software engineering (SE). We also show that many of the items present in the framework are not considered by existing guidelines and checklists.

The development of the framework was motivated by the hypothesis that there is limited existing advice on recruiting *credible participants* (we define this term in Section 1.2) for field studies in SE. Guidelines and checklists in SE research tend to discuss participants in terms of populations, samples and items of interest, e.g., the researcher should define the population of interest and sample items from that population, or from a sampling frame (cf. [1]). Each practitioner recruited is simply an item. But an empirical study needs practitioners who can provide rigorous and valid information that is relevant to that study; in other words, the study needs *credible participants*. Sampling items does not, in itself, ensure or assure that the "right" practitioners are recruited. Also, researchers tend to use a *convenience sample* [2], often with self–selecting participants, e.g., respondents aware of and willing to participate in a survey.

Our impressions of existing guidelines and checklists, our own experience, and our authorship of books on case study research [3] and controlled experiments [4] all motivate our hypothesis and our development of the framework. Furthermore, the framework complements our previous research [5], in which we reasoned about *credible evidence* (defined in Section 1.2), but did not address the credibility of information obtained from participants in different empirical studies in SE.

1.1. Objectives

This article has the following objectives:

1. To review related work to identify material that could contribute to a framework;
2. To review existing guidelines and checklists, and other implicit guidance, to establish whether the recruitment of *credible participants* is considered by such guidance;
3. To formulate a framework for recruiting credible practitioners; and
4. To identify exemplar articles with which we can illustrate the framework.

1.2. Scope of the article, and definitions

To delimit our framework, we focus on the use of participants as providers of information in *field studies*. There is no agreed definition of “field studies” in the SE literature. We therefore define a “field study” as a study conducted with software practitioners that draws on their professional knowledge of aspects of practice, and which takes place in a non–controlled environment. Field studies include case study, action research, ethnography, direct observation, participant observation, interview and survey.

In a previous article [5], we defined “credible evidence” as “. . . an overarching quality aspect of the output of research. . . ” which is built up from a combination of validity and relevance. Drawing on that definition, we define a “credible participant” as a person that we trust, believe or rely upon to provide valid information that is relevant to a researcher.

Furthermore, we focus on the participant as a provider of information about software phenomena that is *external* to the participant. External events are activities that occur in the real world setting, outside of the participant, which the participant experiences, e.g., testing code. Conversely, we *exclude* empirical studies investigating the internal characteristics of a participant, i.e., activities that occur within the participant, e.g. their personal attitude to some event, their motivation, (un)happiness [6], or stress. These internal characteristics may affect the ability of participant to be a credible performer or observer of external behaviour.

We focus our framework on case studies, interviews and surveys of software development as these appear to be the most commonly conducted types of empirical study in SE. We exclude empirical studies of controlled situations, the prototypical example being the experiment. Lenarduzzi et al. [7] are developing an emerging methodology for the selection of participants in software engineering *experiments*. We consider their study later in this article.

1.3. Contribution

The main contribution of this article is to encourage researchers to *think differently* about the involvement of practitioners in field studies of SE. In many field studies, practitioners are not items of interest, to be treated as datapoints or variables, but are, instead, “research instruments” through which, or with which, we gather information about the items of interest. In support of this main contribution we:

1. Corroborate our hypothesis (i.e., that there is limited guidance currently available on the recruitment of credible practitioners) and demonstrate a gap in the empirical software engineering research community's thinking about recruiting participants for field studies. We do this through the analysis of existing guidance, and through the illustration of the framework with three exemplar studies.
2. Propose a framework to help researchers think differently about how to recruit credible participants when collecting data for their field studies. The framework combines several contributions from prior work. Taken individually, none of the prior contributions address the gap/s we identify in existing guidelines and checklists. We synthesis these items of prior work into a framework that takes a different perspective to existing guidelines and checklists, and that makes explicit some issues that have been implicitly recognised in exemplar studies.

1.4. Structure of the article

The remainder of this article is organised as follows: Section 2 reviews background work; Section 3 explains our research approach for reviewing existing guidance, formulating the framework and illustrating the framework with three exemplar studies; Section 4 analyses related work in terms of the guidelines and other advisory sources concerning their treatment of participant recruitment for field studies; Section 5 presents the framework itself; Section 6 illustrates the framework with an exemplar study, a case study; and, finally, Section 7 concludes. Two appendices complement the main article. Appendix A provides a more detailed summary of advice from existing guidelines, to complement Section 4 and Appendix B presents two further illustrations to complement Section 6.

2. Background

Our review focuses on prior work that will contribute as "building blocks", or components, to the framework developed later in this article. We begin with a discussion of the flow of information through an SE research study. This discussion provides a context for thinking about the credibility of participants as providers of information into that flow. We then consider the problem of participant accuracy, referring to two articles [8, 9] published in anthropology. We consider participant accuracy because it can be used as an indicator of participant credibility and because the two articles demonstrate significant challenges with participant accuracy. Next, we discuss sampling in SE research. We distinguish between the item of interest in a sample and the participant as a research instrument to study that item of interest. We use a published field study [10] as an example to illustrate the difference between items of interest and participants. Because participants have a different status, as research instruments rather than items in a sample, we then consider participants as key informants. We connect our preceding discussions to the R^3 model [11] of participant experience, and finally summarise our review of background work.

2.1. The flow of information in SE research

In their study of two large, independent software projects, Karlström and Runeson [12] present a model of the flow of information in the research process. A simplified version of Karlström and Runeson's [12] model is presented in Figure 1. Information about the world is based on the participants' perceptions of the actual world. The information is transformed (cf. T_A and T_B, in Figure 1) as it flows through the research process. This transformation may affect, amongst other qualities, the validity, reliability and relevance of the information.

Because Karlström and Runeson [12] conduct a case study (of two cases), the participants are recruited in relation to specific, identifiable situations in the world. In Karlström and Runeson's [12] study, the *relationship* of the participants to the set of events in the world is therefore relatively well known. A consequence is that the researchers can have more confidence in the information provided by the participants, e.g., because the researchers can assess the relationship. Furthermore, Karlström and Runeson [12] also interview the participants, providing the opportunity (at least in principle) to clarify or challenge the information provided by those participants. In other words, the researchers know, or can know, something about the source of the information and about the nature of the transformations between levels of information (at T_A and T_B, in Figure 1), such as the information–selection decisions being made at levels 2 and 3.

Circumstances can be very different for interview studies and survey studies. The researcher may have less influence or control on the participants recruited, e.g. using a convenience sample. Because of the nature of the interview study, the researcher may know something about the nature of transformations, may be able to influence those transformations (e.g., to gather information based on actual experience rather than cultural norms),

Figure 1: The source and flow of information in research (derived from Karlström and Runeson [12]).

and may be able to know something about the information–selection occurring at levels 2 and 3. But because of the nature of the survey study, the researcher has limited, if any, influence or control on the situations in the world to which the participants refer. The researcher has limited opportunity to influence the information provided by the participants, but nevertheless still has some opportunity. The researcher knows little about the nature of transformations at T_A and T_B, or about the information selected at level 2 and the information shared from level 2 to level 3.

The quality of information flowing through the research is therefore fundamentally dependent on the quality of information provided at the source of the process, i.e., the information provided by the participant.

2.2. The problem of participant accuracy

Bernard et al. [8] consider the problem of participant accuracy (i.e., the degree to which the information provided by the participant conforms to the correct value) and the validity of retrospective data. They observe that anthropology researchers often ask participants to provide data on, as examples, their (i.e., the participant's) behaviour, on the behaviour of others, on sequences of events, and on economic and environmental conditions. Bernard et al. [8] identify three areas in which participant accuracy had been moderately well studied - i.e., recall of childcare behaviour, recall of health seeking behaviour, and recall of communication and social interaction - together with a fourth area that attempted to deal constructively with the problem of participant accuracy.

Bernard et al. [8] present many examples, arguments and conclusions. For conciseness, we present one example here and use that example as the basis for our discussion of their work. The example is drawn from an experiment conducted by Kronenfield et al. [9] in which informants leaving a restaurant were asked to report on what the waiters and waitresses were wearing, as well as the music being played. Kronenfield et al. [9] found much higher agreement about what *waiters* were wearing than what *waitresses* were wearing. This was despite the fact that none of the restaurants in question had waiters. Similarly, they found that informants provided greater detail about the kind of music that was playing in restaurants that were, in fact, not playing music. This example, together with Kronenfield et al.'s [9] interpretation of the results and other work that Bernard et al. reviewed, leads Bernard et al. [8] to suggest the following:

1. Participants who have *actually* observed an event or circumstance are able to report *parts* of the actual event or circumstance. They can only report parts because, for example, they can only recall part of the actual experience.
2. By contrast, those who have *not* observed an event or circumstance start from cultural norms. They are able to provide "...rich descriptions, unencumbered by partial memories and working from complex normative wholes, based on many experiences over a lifetime." ([8], p. 510). In other words, participants who have not actually experienced the event, infer (not necessarily consciously) apparently more complete information about the event.
3. Interviewing many inaccurate participants will not solve the accuracy/validity problem, and will, as a consequence, also not produce relevant findings.

Bernard et al.'s [8] observation about cultural norms is demonstrably present in software engineering. For example, Rainer et al. [13] found that practitioners prefer local opinion over other sources of knowledge. And Devanbu et al. [14] conducted a large survey of Microsoft employees ($n = 564$)

finding that developers' beliefs are based primarily on their personal experience and then, second, on their peers' opinions. By contrast, research articles were ranked fifth out of six. Thus, when practitioners cannot rely on their own experience they appear to first turn to others, i.e., to the source of cultural norms. This raises a serious implication for SE research, i.e., consistency of responses across a *sample* of apparently independent participants may be explained by cultural norms rather than by a consistent behaviour in the phenomenon of interest. In other words, a sample of practitioners may simply provide a representative sample of cultural norms and not provide insights into the phenomenon of interest.

2.3. Sampling

In a recent article, Baltes and Ralph [1] provide a primer on sampling in SE research. They define sampling as the process of selecting a smaller group of items to study, the *sample*, from a larger group of items of interest, the *population*. A *sampling frame* is the available population list from which a sample can be actually drawn.

Baltes and Ralph [1] identify two problems with sampling frames: first, for many software engineering phenomena there is no suitable sampling frame from which to draw a sample; second, some software engineering studies adopt poorly understood sampling strategies such as random sampling from a non–representative surrogate population. Baltes and Ralph [1] also write, "For our purposes, *representativeness* is the degree to which a sample's properties (of interest) resemble those of a target population." ([1]; emphasis in original). Bouraffa and Maalej [2] explore the issues of sampling in more detail, through a review of 54 studies from 41 publications. They found the most frequently used sampling strategy was convenience sampling, with the majority of studies using a reduced sample size of participants drawn from a single organisation. This results in high sample homogeneity, which connects to Bernard et al.'s [8] concerns about cultural norms: a high sample homogeneity is more likely to be sensitive to cultural norms.

As a contrasting perspective, we can *model* a participant as a kind of research instrument - a lens - with which, or through which, we can study software practice. For example, in their field study of software design for large systems, Curtis et al. [10] interviewed 97 participants across 17 projects in 9 companies. In our terminology, Curtis et al. [10] used 97 research instruments to observe the behaviour of software development at five levels of behaviour:

the individual, the team, the project, the company and the business milieu. To clarify, the individual level did not refer to the 97 participants each introspectively studying themselves, but rather the 97 participants provided information on the behaviour at the individual level in software development.

When modelling the participant as a research instrument, the participant is not the item of interest, but is instead a means to study the item of interest. Taking this perspective, the item of interest in Curtis et al.'s [10] study is the software project designing a large system. The 17 projects in 9 companies are therefore the sample of items of interest, and the 97 participants are instruments to study that sample. Furthermore, at least in principle, these 97 participants are selected as the more credible participants for providing information on the 17 items of interest.

2.4. Key informants

As noted in the preceding subsection, the 97 participants in Curtis et al.'s [10] field study were, at least in principle, the more credible participants for providing information. Marshall [15] defines a key informant as an expert source of information. The principal of a key informant is that the informant can provide more reliable, valid and relevant information than a sample of participants. In other words, a key informant can be a more credible participant than a sample. This relates back to Bernard et al.'s [8] review, e.g., that more participants will not in itself solve the accuracy problem, that researchers seek participants who can provide information on the basis of actual experience, and that we seek to avoid information based on cultural norms.

Marshall [15] identifies five characteristics of the ideal key informant, summarised here in Table 1. In principle, all five characteristics contribute to participant credibility. In his article, Marshall simply summarised the characteristics; he did not explain whether or how these criteria are complete in their coverage, or how they might be studied empirically. Furthermore, he recognised that of these five criteria, only the informant's role in the community can be determined with certainty in advance.

2.5. The R^3 model

Falessi et al. [11] propose the R^3 model comprising three elements of a participant's experience: Real, Relevant and Recent. The R^3 model was formulated in relation to participant's *experience* for *experiments*. The R^3

Table 1: Characteristics of ideal key informant (from [15]).

#	Description
C1	**Role in community**. Their professional role in their peer community should expose them to the kind of information being sought by the researcher.
C2	**Knowledge**. In addition to having access to the information desired, the informant should have absorbed the information meaningfully.
C3	**Willingness**. The informant should be willing to communicate their knowledge to the interviewer and to cooperate as fully as possible.
C4	**Communicability**. They should be able to communicate their knowledge in a manner that is intelligible to the interviewer.
C5	**Impartiality**. The key informant should be objective and unbiased. Any relevant biases should be known by the interviewer, e.g., the key informant declares a bias or the interviewer can determine this from other sources.

model is relevant to our work because experience is likely to be an element of practitioner credibility in non–experimental field studies.

We summarise the R^3 model here and then reformulate it in Section 5 to align with our proposed framework.

The three elements of the R^3 model are:

- That the subject has *real experience* of software engineering situations. It is not possible to provide a formal definition for "real experience". Broadly speaking, "real experience" refers to experience of situations of real–world software practice that are, in general, of interest to the research.

- That the subject has *relevant experience*. Relevance here refers to the fit between the situation and the research objective. The characteristic of relevance becomes more significant for a type of informant we discuss later in this article, i.e., the *advisor*.

- That the subject has *recent experience*, or more precisely *timely experience*, e.g., typically that the situation has been experienced recently by the participant, relative to the focus of the research.

All three of the above elements need to be tailored, by the researcher, to the specific needs of the respective research. The researcher will also need to assess each participant against each of the (possibly tailored) elements.

2.6. Summary

We need higher–quality (i.e., more credible) participants in SE field studies, and we also need more participants for those field studies. Treating the recruitment of participants as a matter of *sampling* is potentially limiting our perspective on the problem, e.g., because a sample (however that is defined) may simply be representing cultural norms. In this section, we have reviewed prior research to identify several insights into the recruitment of participants. These insights act as inputs, i.e., as building blocks, into the development of our framework, discussed in Section 5.

3. Research approach

This section explains how we reviewed existing guidance on the recruitment of participants for empirical studies, how we developed the framework, and how we illustrated the framework using exemplars.

3.1. Reviewing guidelines and other advisory sources

To identify appropriate related work in terms of guidelines and other sources of advice, the two authors independently searched for appropriate articles, primarily using Google Scholar. We use Google Scholar because we seek a sufficient coverage of guidelines and not an exhaustive coverage; we are not, for example, attempting a systematic review. We also, of course, had prior experience of some guidelines, e.g., the case study guidelines by Runeson and Höst [16]. We shared the suggested guidelines, and discussed them in online meetings, arriving at consensus on the advisory sources to consider.

We prioritised guidelines and recommendations that focused on field studies, or empirical studies in general, and that were published *after* 2009, to be consistent with when the case study guidelines were published. Thus we excluded guidelines, such as Lenarduzzi et al.'s [7], that focused on experimental studies. We had difficulties finding guidelines on interviews, discussed below.

In total we identified six initial sources. These are listed in the upper part of Table 2 and briefly summarised as follows.

Case study: Research guidelines for conducting case study research within software engineering were first published as a checklist, by Höst and Runeson [17], at a conference in 2007. Runeson and Höst then published a more extensive set of guidelines as a journal article in 2009 [16] and then as a book [3] in 2012. Verner et al. [18] published guidelines for industrial case studies in software engineering at a conference in 2009.

Interview study: We did not find guidelines for interview studies in software engineering that are comparable to the case study guidelines or to the survey study guidelines (discussed next). We therefore had to relax our requirement for post–2009 publication for these interview "guidelines". We also chose two articles. The first article, by Strandberg [19], provides advice concerning ethics in interview studies. The second article, by Hove and Anda [20], shares their experiences of conducting interview studies in software engineering.

Survey: Empirically evaluated survey guidelines are published by Molléri et al. [21] in 2020. Their article synthesises a range of previous guidelines on survey studies, e.g., [22, 23].

Participant recruitment: Salleh et al.'s [24] article does not present *guidelines* but rather, in a way that is similar to Hove and Anda [20], Salleh et al. [24] share their experiences of conducting research in industrial contexts.

We wanted to go beyond just the guidelines for case studies, interview studies and survey studies to consider guidelines for other types of field study. Consequently, after we identified these initial sources, we subsequently also then identified four additional sources for reviewing the selection of participants in different types of field study. The four additional sources are listed in the lower part of Table 2. Again, we needed to relax our requirement for post–2009 publication. The four additional advisory sources are:

Focus groups: Kontio et al.'s [25] book chapter, published in 2008, provides guidelines on the use of focus groups in SE.

Preliminary guidelines: Kitchenham et al. [26] published probably the first set of guidelines on empirical studies in SE, in 2002.

Ethnography: Zhang et al. [27] published guidelines for ethnographic studies in 2019.

Empirical standards: Ralph, in conjunction with ACM SIGSOFT, are developing standards for empirical studies in SE [28] . These standards were first published in 2021.

For all ten articles listed in Table 2, we downloaded PDF copies of each article and searched each PDF for explicit guidance on participant selection. The objective of the searches was to identify formulations in the guidelines and advisory sources in relation to advice concerning recruitment of participants for field studies. We were particularly looking for concrete advice, i.e., beyond general statements concerning the importance of recruiting representative participants, e.g., through sampling. To do this, we used the search facility of our PDF viewer to automatically search for stemmed words related to the persons in a study and to the activity of recruiting such persons. We searched for the following nine stemmed words:

For person: `subject*`, `partici*`, `respond*` and `contribu*`

For activity: `select*`, `identif*`, `sampl*`, `find*` and `recruit*`

We discuss our analysis of the ten articles in Section 4.

Table 2: Publications selected for reviewing advice on participant recruitment.

Ref	**Year**	**Description**
Review existing advisory sources		
[16]	2009	Case study guidelines by Runeson & Höst
[18]	2009	Case study guidelines by Verner et al.
[19]	2019	Ethical interviews
[20]	2005	Semi-structured interviews
[21]	2020	Survey guidelines
[24]	2018	Recruiting participants
Further review of advisory sources		
[25]	2008	Focus group guidelines
[26]	2002	Empirical studies guidelines
[27]	2019	Ethnographic guidelines
[28]	2020	Empirical standards

3.2. Development of the framework

We used a dialectic process to develop the framework, i.e., the first author devised, and subsequently revised, the framework, and the second author independently reviewed the latest version, providing feedback which lead to subsequent revisions. The dialect process was informed by our respective prior research experience, particularly the books [3, 4], as well as the items identified in Section 2, e.g., information flow, participant accuracy, the characteristics of key informants, and the R^3 model [11]. Being highly–cited, a number of these sources – specifically, Bernard et al., [8], Marshall [15], and our two co-authored books [3, 4] – are established in their respective fields of research. The nature of the dialectic process means that the framework is not *deduced from* these sources and items but rather *created with* them. The review of guidelines in Section 4, and the illustration of the framework with exemplars in Section 6 and Appendix B, show there is the need to think more carefully about these issues.

Furthermore, the framework progressed through five revisions before we moved to our illustration of the framework. After completing our illustrations, we returned to refine the *presentation* of the framework, i.e., we did not change the content of the framework but simplified the way in which it is presented and described.

3.3. Illustrating the framework with exemplar articles

Having developed the framework, we wanted to confirm whether the components of the framework can be found in at least some published field studies. Our objective here is not to assess the prevalence of the framework's components in prior research; rather, we want to simply illustrate that these components are considered relevant in at least some of the articles that have good descriptions of participants and their recruitment.

To select primary studies for illustration, both authors independently searched for candidate articles to consider. We used the following search heuristics:

- We used Google Scholar for the searches.

- We prioritised the more highly cited articles, for two reasons: first, we assumed that the more highly cited articles were more likely to have valuable information in them; second, that the research community would have a greater awareness of these articles and therefore our illustration would have more obvious relevance to the community.

- We prioritised articles having better descriptions of the participants in the studies. Better descriptions would help us more easily find information about the framework's components. By contrast, in an article that reported less information about participants, the study itself may have considered the components, but simply not reported them.

We then independently read the list of candidate articles and discussed them in online meetings, and agreed on the final selection. During our final selection we considered the following criteria:

- Articles published before 2009, since 2009 is the year that both Runeson and Höst [16] and Verner et al. [18] published their guidelines on case study research in SE. We chose articles published prior to 2009 since such articles ought to have influenced the formulations of the guidelines.
- Articles reporting primary studies.
- Articles that, taken together, would provide coverage of case study, interview study and survey study.
- Articles that, taken together, would provide coverage across journals and conferences.

Table 3 lists the three articles we selected. We discuss these articles in Section 6 and in Appendix B. Table 3 reports only the articles finally selected for the illustrations. Many other articles were considered.

Table 3: Publications selected for illustrating the framework.

Ref	**Year**	**Description**
Assess framework against primary studies		
[29]	2005	Case study
[30]	1998	Interview
[31]	2002	Survey

4. Analysis of published guidelines and other advisory sources

As discussed in Section 3.1, we searched the PDF files of the ten guidelines and advisory sources. The outcome is summarised in Table 4. The table includes the number of occurrences of the stemmed words listed in Section 3.1, together with the number of relevant sentences found in the respective article, and the frequency of those relevant sentences in relation to the total number of occurrences of the stemmed words. We summarise the advice given in the ten articles, in Tables A.1 and A.2 in Appendix A. These summaries provide a sense of the focus and coverage of the guidelines, i.e., that existing guidelines are quite "light" in their coverage of participant recruitment.

Table 4: Summary counts of searches of stemmed words.

Article	**Research method**	**Occurrence** f	**Quote** f	**Quote** %
Initial selection of six advisory sources				
Runeson [16]	Case study	125	8	7
Verner [18]	Case study	87	8	9
Strandberg [19]	Interviews	101	3	3
Hove [20]	Interviews	78	4	5
Molléri [21]	Survey	312	6	2
Salleh [24]	General	293	8	3
Further selection of four advisory sources				
Kontio [25]	Focus group	138	8	6
Kitchenham [26]	General	114	4	4
Zhang [27]	Ethnography	116	1	1
Ralph [28]	Standards	45	1	2

To complement Table 4 and Appendix A we provide brief summaries of the advisory sources below.

The guidelines by Runeson and Höst [16] highlight the importance of recruiting suitable participants in relation to the objective of the case study, however the guidelines do not provide guidance on how to make an informed decision concerning participant recruitment. Verner et al. [18] provide case study guidelines for, as they call it, industry–based studies in software engineering. When it comes to selecting participants, they touch on the subject when providing an example concerning the scope of the case study. However, the guidelines by Verner et al. do not further address selecting credible participants except for mentioning the importance of determining the sampling strategy.

Strandberg [19] highlights the need to know who the stakeholders are, and to be able to consider the potential benefit and harm that may arise from the research. Strandberg does not, however, discuss how to identify appropriate participants as interviewees.

Hove and Anda [20] highlight the need to select or recruit participants carefully. However, Hove and Anda do not provide experiences concerning the challenges of selecting suitable participants in an interview study.

In the survey guidelines, Molléri et al. [21] stress that we should identify and select participants based on characteristics, however their article does not provide support concerning what constitutes essential characteristics when recruiting participants for a survey.

Salleh et al. [24] highlight that it is essential that specific requirements on the participants need to be conveyed to the industrial collaborator. However, Salleh et al. [24] do not provide further details concerning what may make a participant suitable for participation in the research.

Kontio et al. [25] suggest using purposive sampling, i.e., participants are selected based on their characteristics in relation to the topic of the focus group session.

Kitchenham et al. [26] provide more generic guidance on empirical studies, though their advice is more focused on controlled experiments and statistical analysis. They argue that subjects should be representative of the population. Their preliminary guidelines do not, however, discuss any specific desirable characteristics concerning the participants.

Turning to more recent guidelines, Zhang et al. [27] present a critical review and checklist for conducting ethnographic studies in software engineering. The guidelines do not provide guidance on how to assess the credibility of the practitioners being studied.

Finally, the empirical standards [28] only mention participants in relation to quantitative studies, and do not discuss participants in more qualitative studies such as addressed here.

Overall, none of the ten guidelines and other advisory sources provide actionable advice on *how* to determine the credibility of prospective participants.

5. Formulating a framework about credible practitioners

In this section, we present and discuss our framework for thinking about credible participants and the quality of information they can provide to a field study. The framework uses the "building blocks" introduced in Section 2. We first introduce and discuss several components of the framework, and then concisely present the framework in Section 5.6 and Table 5. We briefly described, in Section 3.2, how the framework was developed.

For the components, we begin with a simple model of the research process as a reference. We then re-consider the sampling of participants for empirical studies, re-framing this as a problem of recruiting credible participants. Then we introduce the components of the framework – i.e., the three participant roles, characteristics affecting the quality of information, and demographics – before concisely presenting the framework in Section 5.6.

Following the presentation of the framework, we describe additional considerations that are beyond the scope of the current article, present a simple example of the application of the framework, and summarise the contribution of the framework. In Section 6 and Appendix B we illustrate the framework.

5.1. A model of the research process

Figure 2 presents a simple model of the research process for field studies in SE. The model is intended to be used as a reference for the subsequent discussion. In the model, a theory of some kind provides the grounds for a proposition. The proposition is studied empirically.

As already noted, the reference model is a simplification. For example, grounded theories are generated bottom–up from the empirical world. As another example, the propositions of the model may be hypotheses, research questions, or other kinds of testable or empirically investigatable statements.

For empirical SE research, many aspects of SE practice can be studied directly, e.g., source code, however many other aspects of software engineering practice can only be studied indirectly, e.g., through engaging with software practitioners who themselves interact with the empirical world of SE. In the model, practitioners provide information about the empirical world to the researchers as part of an empirical study. As discussed in Section 2, these practitioners may therefore be understood as research instruments. The information that practitioners provide to researchers is broadly of two types: facts that *describe* some aspect of a *specific* software engineering situation, and beliefs about practice that may be specific to a situation or be generalised to more than one situation. Participant demographics may be understood as factual information, however our focus here is on both the factual information and the beliefs that participants provide about the phenomenon of interest.

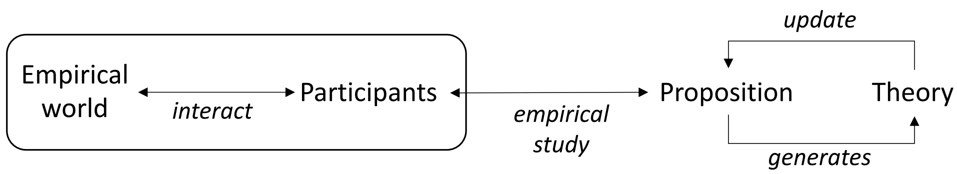

Figure 2: A simple reference model of the research process for field studies in SE.

Again for simplicity, we assume that theories are constructed from four fundamental constructs: the *actor*, the *technology*, the *activity* and the *software artefact*. These constructs are well–accepted in software engineering research [32]. These constructs exist within a *context* [33, 34, 35] which, by its nature, is difficult to define.

The scope of the theory, and therefore of the propositions, will align in some way with the empirical world being field–studied. For example, a researcher working with a theory and propositions about requirements engineering is unlikely to be empirically investigating code inspections.

Participants in a field study are expected to be drawn from the empirical world as that world aligns with the theory and propositions. Remaining with our example, the researcher would likely conduct field studies with requirements engineers as the participants, and ask those requirements engineers about requirements engineering. Also, the researcher will empirically study attributes of the four constructs as they relate to requirements engineering, e.g., requirements engineers (*actor*), who use requirements gathering templates (*technology*), to elicit (*activity*) requirements (*software artefact*), all within a *context*.

5.2. Participants as a sampling frame

In Section 2.3, we discussed the sampling of items from a population of interest. We suggest that, depending on the theory and propositions, it may be more effective to treat participants as a kind of sampling frame through which the items of interest from the empirical world are sampled, and therefore indirectly studied. Our suggestion is illustrated in Figure 3.

Remaining with our earlier example, if the researcher intends to study the *attitude* of requirements engineers then it makes sense to treat requirements engineers as the population and to sample from that population. This is because attitude is a property, or attribute, of the requirements engineers themselves. But if the researcher intends to study any one or more of the

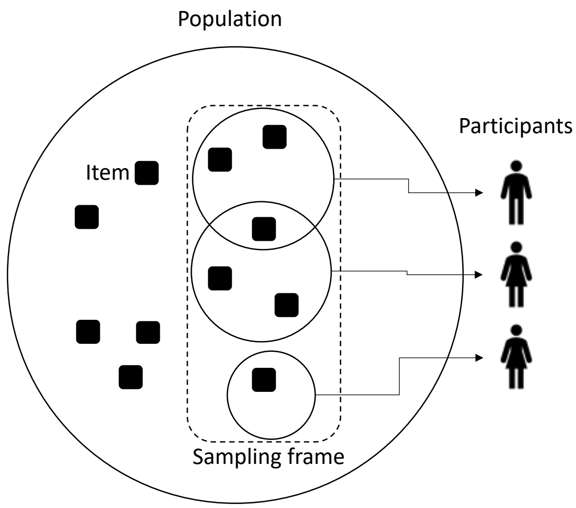

Figure 3: Participants as a sampling frame. Each participant has access to one or more items of interest from the population, and the aggregate of participants provides a sampling frame for sampling the items of interest.

other three constructs of theories – i.e., *activity*, *technology* or *artefact* – or *actors* other than the participant, then the population of interest is not the requirements engineer but rather one or more of these other constructs. More strictly, the population of interest is likely to be a configuration of *actor*, *activity*, *technology* and *artefact*, all within one or more contexts. The researcher should ideally sample from across all of the appropriate constructs of interest. The implication is that participants should be recruited for the "access" they give to the population of *actor*, *activity*, *technology* and *artefact*, and not sampled for their representativeness as practitioners. In other words, a participant should be recruited for the contribution they can make to the formation of a sampling frame for sampling the items of actual interest. Recruiting participants in this way helps to ensure that the researcher collects information – either facts about the world, or beliefs about the world – that are drawn in relation to the items of interest.

5.3. Participant roles

Participants will often be software practitioners who are located somewhere *within* a software engineering situation and can therefore perceive other *actors*, as well as the *activities*, *technologies* and *artefacts* of the situation. Using Falessi et al.'s [11] R^3 model, discussed in Section 2.5, practitioners would be expected to have some degree of Real, Relevant, and Recent experience.

Sometimes a participant may be a software practitioner who is located outside of the situation but who can, for some reason, contribute to the formation of a sampling frame. Also, sometimes the practitioner has real and relevant experience, but the experience is not recent.

For our framework, we therefore define three roles for a participant in a field study:

- The participant who is a *Performer* within the situation, e.g., a programmer.
- The participant who is an *Observer*, but not a *Performer*, and is located elsewhere within the situation, e.g., a tester may observe aspects of the programmer's behaviour and performance.
- The participant who is an *Advisor* with experience from a range of other, but related, situations. A common example here is a consultant who has not performed in the particular situation, or observed it, but draws on professional experience from elsewhere. The more experienced, and the more widely experienced, a software practitioner becomes the more likely they will have *some* real, relevant experience, though that experience may not be recent.

Practitioners in each of these roles provide information to the researcher, but this information is of different degrees of credibility, e.g., the information may be drawn from real experience but the experience, and therefore the information, may not actually be relevant. Similarly, the *Performer* will have experienced a situation contemporaneously (they did something at the time), but that experience may no longer be recent in relation to when the field study is being conducted. Later in this section, we map the three roles to the three elements of Falessi et al.'s [11] R^3 model.

A practitioner's experience may allow that practitioner to be classified as a *Performer* for one study, whilst some other of the same practitioner's experience may allow that practitioner to be classified as an *Observer* for another study, and some other of the same practitioner's experience may allow that practitioner to be classified as an *Advisor* for yet another study; and indeed some of the practitioner's experience may not allow classification according to the three roles, e.g., because that experience is not relevant.

For example, a software tester may be a *Performer* in a study of software testing, an *Observer* in study of programming, and an *Advisor* in a study of testing in another context. We acknowledge the complication of particular practitioners having multiple roles; it is a complication inherent with research into SE.

We also recognise the potential role of an *expert* and intentionally do not use the word "expert" as a label for any of our roles. This is because a *Performer* or an *Observer* or an *Advisor* may be an expert, e.g., depending on their experience.

Our suggestion is that, when undertaking an empirical study, and when *recruiting* participants, the researcher evaluates a practitioner against these three roles in the order that we have presented them, i.e., first determine whether the practitioner could be treated as a *Performer*. According to our framework, a practitioner who could not satisfy any of the three roles would in principle be rejected as a participant in the study. In practice, the researcher may introduce an additional role or roles for practitioners, where it is appropriate to do so for their research. For example, researchers may introduce the role of *Client* in a requirements engineering activity to recognise a participant who is able to make observations of the requirements engineer, but perhaps lacks relevant experience of requirements engineering to be an *Observer*.

5.4. Practitioner characteristics affecting the information they provide

In addition to the three participant roles, we suggest five characteristics of the practitioner that might affect the quality of information they can provide as a participant in a study. We identified these characteristics using the dialectic process briefly described in Section 3.2. The five characteristics are:

Quantity of experience: In general, the greater the quantity of situations experienced by the practitioner, the more experienced is the practitioner for the research.

Perceptual sensitivity: The more the practitioner is able to *carefully* perceive events and, through that careful perception, to help prevent or reduce bias in their perception, the more valuable is the practitioner for the research.

Situation selectivity: The more the practitioner can distinguish between the different situations and so provide only the more real, relevant and recent information to the researcher, the more valuable is the practitioner for the research. For this characteristic, it may be more valuable to the researcher for the practitioner to *restrict* the information they share to a specific situation, or situations, to ensure the information is more valid and more relevant, cf. the role of the sampling frame.

Reflexivity: The more the practitioner is able to subsequently *reflect* on those perceptions and, through that reflection, to help prevent or reduce biases in the information they share with the researchers, the more valuable is the practitioner for the research.

Willingness: The practitioner must have the willingness or, alternatively phrased, the motivation to openly share information.

These five characteristics are all simplifications, and are all very challenging to measure. Again, the aim is to encourage researchers to think carefully about the recruitment of more credible participants.

5.5. Demographics of participants

For clarity, we distinguish between the demographics of participants and what might be called the demographics of the phenomenon of interest, e.g., the context of the *actor–activity–technology–artefact* configuration. Given our focus on recruiting credible participants, this subsection focuses on *participant* demographics. In Section 6 we also recognise the demographics, or context, of the phenomenon of interest.

Empirical studies of software engineering often report the demographics of the participants. For example, in one of their survey studies of professional software engineers at Microsoft, Begel and Zimmerman [36] found statistically significant differences in responses to 29 questions, based on the demographics of the respondents. And taking one pervasive demographic – gender – as another example, Carver and Serebrenik [37] summarise several papers published in the 2019 edition of the International Conference on Software Engineering (ICSE) that discuss gender and software engineering.

These examples highlight the importance of considering demographics when collecting information from practitioners about their experience of software practice; practitioners with differing demographics may have differing experiences and may therefore provide different information. Also, it may be that the three roles we have identified have different demographic profiles, e.g., different proportions of gender or age for the different roles. Because of the inherent nature of demographics, and its relevance to software engineering research and practice, we explicitly recognise demographics in our framework, rather than only implying it in the existing characteristics.

It is essential to note that we do not view experience as a demographic; it is one of the characteristics listed by Marshall. In our opinion, it is better to handle experience as a characteristic since it relates to the individual as an information provider, while the demographics provide general descriptive information about the information provider.

Also, for clarification, we include the practitioner's *functional role* (e.g., Systems Engineer, Programmer, or Tester) as part of the participant's demographics. We include functional role under demographics for a few reasons. We want to include functional role in the framework whilst also recognising that both functional roles and the *titles* for functional roles can vary substantially from project to project, and from company to company, etc. For these reasons, functional role is not necessarily a reliable indicator of the credible practitioner and is therefore not included in our five quality characteristics.

5.6. A concise presentation of the framework

A concise summary of the framework is presented in Table 5. As noted earlier, participants can provide two types of information to researchers, i.e., facts about the phenomenon of interest and beliefs about the phenomenon. The table organises the participant roles, the R^3 model, the five characteristics, the demographics, and the two types of information. Different roles can provide different types of information and, depending on the characteristics of the participant, different degrees of quality of information. We intentionally do not try to specify the details of the Quality criteria; these are complex criteria that are hard to measure and our primary interest in the current article is to encourage researchers to *think about* these criteria when recruiting practitioners.

5.7. Additional considerations

When applying the framework some additional considerations will need to be addressed. As one example, focusing on the more credible participants might reduce the number of participants and therefore increase the likelihood that a given participant might be identifiable, e.g., when conducting an interview study in a company. This raises ethical issues. We do not consider these

Table 5: A summary of the framework.

Role	R^3 model: Real	Rlvnt	Rcnt	Quality: E	P	S	R	W	D	Information: Fact	Blfs
Performer	YES	YES	YES	...					...	YES	YES
Observer	YES	yes	yes	...					...	yes	yes
Advisor	yes	maybe	maybe	...					...	–	yes

Notes for R^3 model:

YES = strongly meets criterion; yes=moderately meets criterion

maybe = may meet criterion depending on context.

Notes for Quality criteria:

We intentionally do not populate the Quality criteria.

E = Quantity of experience; P = Perceptual sensitivity;

S = Situation selectivity; R = Reflexivity; W = Willingness

Notes for Demographics (D):

We explicitly recognise demographics because of its relevance to SE.

Notes for Information:

Blfs = Beliefs

YES = more likely to provide credible information

yes = may be able to provide credible information

– = cannot provide facts about the item/s of interest issues in this article, however Strandberg [19], as one example, discusses ethical issues affecting interviews in software engineering. As a second example, *finding* the more credible participants may be challenging, particularly as the researcher becomes more selective on the characteristics of the participant. But, as we noted from Bernard et al.'s [8] work in Section 2.2, collecting information from many inaccurate participants will not solve the accuracy problem, and will, as a consequence, also not produce relevant findings.

5.8. Illustrating the framework: a brief example

In Section 6 we illustrate the framework in detail with an exemplar study, and complement that illustration with two further illustrations in Appendix B. Here we present a simple example to illustrate aspects of the framework.

Returning to our earlier example, a researcher may invite requirements engineers to participate in a field study. The framework encourages the researcher to consider whether and how the invited practitioners provide access to the empirical world, as well as the quality of the information the participant can provide.

A requirements engineer (*actor*) who is actively undertaking requirements engineering (*activity*) using appropriate resources (*technology*) and producing requirements specifications (*artefacts*) that align with the theory and propositions of the research model is expected to have the most direct contact with the empirical world; in other words, may be understood as a *Performer* who can, in principle, provide the most credible information to the researcher: facts about the world. Overall, the *Performer's* information is expected to be more Real, Relevant and Recent (timely) compared to, for example, a requirements engineer who is an *Observer*. The least credibility occurs with the requirements engineer who is an *Advisor* as this participant has both the greatest variability in their experiences; but also, the researcher has the *least certainty* in the Realness, Relevance and Recentness of the *Advisor's* experience and information for the respective study.

The three roles provide a convenient way of evaluating practitioners when recruiting participants for a study. These three roles support a coarser–grained evaluation, however; the five characteristics, discussed in sections 5.4, provide a finer–grained approach to evaluating practitioners for recruitment, but are much harder to implement.

By contrast, a common approach to recruiting participants for field studies in software engineering appears to based on the practitioner's functional role (e.g., project manager, software engineer, software tester) and their years of experience.

5.9. General guidelines for using the framework

In Section 1.3, we explained that the framework is intended to help researchers *think differently* about how to recruit credible participants. What counts as a credible participant will depend on the objectives of the specific study. It is therefore not possible to provide *a priori rules* on the recruitment of participants. Instead, we suggest the following general guidelines when using the framework:

1. Where possible, assess *each* participant according to the framework, e.g., the degree to which the participant satisfies each of the components of the R^3 model, and of the Quality criteria. This, of course, requires the development of an assessment scheme.
2. To optimise the recruitment of *Performers* over *Observers* and of *Observers* over *Advisors*.
3. To report the number of participants in each role and, where appropriate, to report measures of the degree to which participants satisfy the components of the framework. Such reporting provides transparency on the credibility of the informants and may support future meta-analyses.
4. To consider collecting and also analysing the information from each of these roles *separately*, and in the order of priority, e.g., for interviews, to collect and analyse information from *Performers* first; for surveys, to select and analyse the *Performers* responses first.
5. To consider using methods of cross-sample comparison or triangulation of data, e.g., comparing information from *Performers* with information from *Observers*.

Finally, we recognise that it may not be possible to recruit the ideal participants for a study, for example it may only be possible to recruit *Advisors*. In such situations, it may still be valuable to conduct a study. Reporting information on participant credibility allows others to assess the quality of the study.

5.10. Contribution of the framework

In this section, we combined contributions from prior research to propose a framework for *thinking about* credible participants. The framework combines the components described in Section 2, i.e., a model of the flow of information through the research process, based on Karlström and Runeson [12]; aspects of participant accuracy, based on Bernard et al. [8]; concepts of sampling, e.g., from Baltes and Ralph [1]; Falessi et al's [11] R^3 model for participant experience; Marshall's [15] five criteria for key informants; and the accepted importance of demographics. Taken individually, none of the prior contributions address the gap/s we identify in existing guidelines and checklists; a synthesis of these prior contributions is required.

Recruiting participants for fields studies has both similarities with, and differences to, recruiting participants for experiments. Indeed, our framework incorporates Falessi et al.'s [11] R^3 model, a model that was originally developed for experiments. In a recent paper, Lenarduzzi et al. [7] present a vision of a methodology for recruiting participants for experiments. They also draw on Falessi et al.'s [11] R^3 model. As well as the R^3 model, their envisioned methodology includes identifying participant characteristics, suitable tools for measuring the characteristics, measuring the distance between the sample and the population, and strategies for minimising such distance.

One distinguishing feature of field studies is the vital importance of the practitioner's knowledge about their professional setting, as recognised in the framework through the roles of *Performer*, *Observer* and *Advisor*. Knowledge of the professional setting – of the *field* – is not typically relevant for an experiment because experiments do not tend to investigate events in the field (if they did, they would more likely be field studies). One implication of this distinguishing feature is that participants in field studies may be good candidates for experiments, where the requirement for knowledge about a professional setting is often not prioritised, but participants in experiments normally do not have the professional knowledge required to participate in a field study. Moreover, participants in field studies ought to have knowledge about the same setting, as in case studies, or similar settings, for example, in a survey on a specific topic. In other words, although the R^3 model may be used for both experiments and, here, for field studies, this does not mean that the same kind of participant can be used for the two types of study.

Knowledge of the field relates to Marshall's [15] characteristics for key informants, such as the *Role in the community* characteristic (see Table 1) which, again, is unlikely to be relevant for an experiment, although other characteristics may be essential. Furthermore, detailed demographics are vital in field studies, and may be less important in experiments. Thus, whilst there are components of the framework that are common to field studies and to experiments, there are also important differences.

6. Illustrating the framework with an exemplar study

6.1. Overview

In this section, we illustrate the framework in detail with one exemplar study, a case study [29]. Two other exemplars, an interview study and a survey study, are considered in detail in Appendix B. In Section 6.3, we summarise all three exemplars in relation to the framework. In Section 3.3, we explained how we selected the exemplars.

Our illustration of the framework is clearly not exhaustive; it does not need to be. It is sufficient to show enough of a mapping from framework to exemplars to illustrate the framework. There are excerpts from the exemplars that we might report here, and in Appendix B, and a given excerpt may be relevant to more than one component of the framework. Furthermore, a mapping of some part of an exemplar into the framework is not an indicator of how substantially the exemplar addressed the respective component, only an indicator that the respective researchers were aware of the component and took some action to address it. Similarly, the researchers of the respective exemplar may not have considered a component to the same level of abstraction that we have in the framework.

6.2. Case study

The case study we consider was published by Freimut et al. [29] as a journal article in 2005. The article proposes a model to measure cost–effectiveness of inspections, as well as a method to determine cost-effectiveness by combining project data and expert opinion. Expert opinion was gathered through interviewing 23 experts. The article is particularly relevant to our framework for two reasons: first, the article discusses the nature of expert opinion; second the article illustrates many components of the framework.

Before discussing the details of Freimut et al.'s article [29] it is helpful to clarify two terms used in the article. First, Freimut et al.'s use of the term "expert" maps most closely to our use of the term "Performer". As briefly discussed in Section 5.3, we consider that *Performers*, *Observers* and *Advisors* may each have a degree of expertise, e.g., that a practitioner might be a relatively novice *Performer* or, alternatively, an expert *Performer*. Second, Freimut et al. [29] use the term "role" to refer to the *functional* role of the practitioner, i.e., they distinguish between practitioners who are Developers, Analysts and Testers. In our framework, by contrast, the *functional* role is a component of participant demographics; we reserve the term "role" in the framework for *Performer*, *Observer* or *Advisor*.

Turning to the details of the article, and first considering the nature of expert opinion, Freimut et al. [29] devote a whole section of their article to discussing the nature of, and the need for, expert opinion. They recognise that one problem, for which expert opinion can help, is when information about a phenomenon cannot be collected by any other viable means, such as through measurement, observation or experimentation. Freimut et al. [29] recognise that expert data is subject to bias, uncertainty and incompleteness, but also that these problems can be controlled through carefully performed elicitation of expert estimates.

Furthermore, Freimut et al. [29] discuss the selection of experts and present two criteria for selection:

1. The *role* of the expert in the development process. For Freimut et al., experts must have access to the information they are supposed to estimate. To do this, the experts must have participated in the respective process, e.g., in design inspections. Freimut et al. write, "...for the purpose of effort estimation, *people performing the corresponding tasks* qualify as experts." ([29], p. 1081; emphasis added).
2. The level of expertise: Freimut et al. write, "Such experience needs to be *sufficiently varied and extensive with respect to the targeted tasks*." ([29], p. 1081; emphasis added).

In terms of illustrating that the Freimut et al. [29] study implicitly uses the components of the framework, we organise the illustration in terms of the main components of the framework:

- **Information:** In the case study, *factual* information was collected as project data from the Quality Assurance management. The experts were asked to provide *beliefs*, e.g., as probability distributions, about defects and costs. Freimut et al. performed several analyses to assess the validity of those beliefs.

- **Quality**: As recognised in Section 5.6 of our article, the Quality criteria are hard to measure etc. There are some indications that Freimut et al. were at least implicitly thinking about some of these criteria. For example:

- **Quantity of experience:** this is indicated by the experts' years of experience.
- **Perceptual sensitivity:** for this criterion, there are no clear examples in the Freimut et al. article.
- **Situation selectivity:** as one example, Freimut et al. discuss how, in the pilot study, experts were accidentally lead to think of unusual instances rather than typical cases.
- **Reflexivity:** as one example, Freimut et al. discuss how an Analyst reflected on how they did not have recent experience. As another example, Freimut et al. discuss how testers in the pilot study unanimously agreed that it was possible for them to estimate the required parameters with their experience, though also recognised that some parameters would be more difficult to estimate.
- **Willingness:** Freimut et al. briefly discuss how experts were recruited on the basis of being proactive and motivated and how the recruited practitioners were "highly motivated during the interviews" and "showed great interest in the study."

- **Demographics**: Freimut et al. do not explicitly discuss demographics; the one indication of demographics in their article is implicitly in terms of gender. When they refer to experts in their study they refer only to male experts, e.g., "Similarly, an expert who is currently not performing the task about which *he* is to provide an estimate may provide different estimates than an expert who is performing the task." ([29], p. 1091; emphasis added).

- **The $\mathbf{R^3}$ model**: Freimut et al. provide a range of comments that clearly map to the R^3 model. As examples:

  - **Real**: The levels of expertise for the experts ranged from 3 to 18 years.
  - **Relevant**: Freimut et al. discuss how a particular expert, who is responsible for a very complex part of the system, provides estimates that are significantly different from other experts who are responsible for less complex parts of the system. All of the experts have Real experience; here, the Relevance of the expert's

opinion is based on the complexity of the part of the system on which the developer is working.

  - **Recent**: Freimut et al. explain that one Analyst had not participated in analysis inspections for a long time and, because the Analyst's estimates differed from the other Analysts, Freimut et al. decided to exclude this estimate.

- **Roles:** Freimut et al. identify three functional roles, i.e., Developer, Analyst and Tester. All three of these functional roles map to the role of *Performer* in the framework. Each functional role *performs* a different activity in software development and is therefore appropriate for providing information on different aspects of the cost–effectiveness of inspections.

6.3. Summary of exemplars

Table 6 and Table 7 summarise the mappings from the three exemplar studies to the framework. Table 6 provides more detail on the mapping, at least for some of the components of the framework, whilst Table 7 provides a concise mapping that directly aligns with Table 5 in Section 5.6. As noted earlier, detailed illustrations using the second two exemplars [30, 31] are presented in Appendix B.

The exemplars include insights on credible participants in field studies in SE that are not covered in the advice offered in the guidelines and other advisory sources. Thus, we argue that the three exemplars show the need for a framework of this kind, and also show that existing guidelines provide limited advice on the topic of recruiting credible practitioners for different forms of field study.

7. Conclusion

We hypothesised that there was limited existing advice on recruiting *credible participants* for field studies in SE. We also argued that *sampling* participants will not, in itself, generate credible evidence as, for example, participants may simply be reporting cultural norms. A review of existing guidelines, checklists and other advisory sources in SE research corroborated our hypothesis.

Table 6: Summary of exemplar articles mapping to the framework.

Criterion	Exemplars: Freimut et al. [29]	Exemplars: Singer [30]	Exemplars: Vredenburg et al. [31]
Meta-information about the study			
Year	2005	1998	2002
Topic	Inspections	Maintenance	User centred design
Study type	Case study	Interview	Survey
Sample	23 people	10 sites	103 people
Medium	Journal	Conference	Conference
Information			
Beliefs	Estimates	"Truths"	Methods etc
Facts	Costs	Background information	Project profiles
Quality criteria			
Experience	Yes	Yes	Yes
Perception	None	None	None
Selection	Some	None	Yes
Reflection	Some	Some	None
Willingness	Yes	Some	Some
Demographics			
Demographics	No	No	Some
Falessi et al.'s [11] R^3 model			
Real	Yes	Yes	Yes
Relevant	Yes	Yes	Yes
Recent	Yes	Yes	Yes
Role			
Performer	Developers (7) Analysts (n/16) Testers (m/16)	Maintainers	Probably
Observer	None	None	Possibly
Advisor	None	None	None

Table 7: Comparison of exemplar articles with the framework.

Art-icle	**Role POA**	**R^3 model**			**Quality**						**Info.**	
		Real	**Rlvnt**	**Rcnt**	**E**	**P**	**S**	**R**	**W**	**D**	**Fact**	**Blfs**
1 [29]	P - -	Y	Y	Y	Y	N	y	y	Y	N	Y	Y
2 [30]	P - -	Y	Y	Y	Y	N	N	y	y	N	y	Y
3 [31]	p o -	Y	Y	Y	Y	N	Y	N	y	y	y	Y

Notes:
Y = There is clear evidence that the article maps, for this component.
N = There is no evidence that the article maps, for this component.
y = There is some evidence that the article maps, for this component.
P = Performer; O = Observer; A = Advisor ; - = None
p = probably Performer; o = possibly Observer
E = Quantity of experience; P = Perceptual sensitivity; S = Situation selectivity
R = Reflexivity; W = Willingness; D = Demographics

Drawing on several contributions from prior research, we proposed a framework for credible participants. The framework is concerned with participants as providers of information to field studies, in which the phenomenon of interest is *external* to the practitioner.

The main contribution of the framework, and of this article, is to encourage researchers to *think differently* about the recruitment and involvement of practitioners in their field studies of SE. Rather than thinking of practitioners as items of interest, we suggest that researchers think of participants as "research instruments", or as sampling frames "into" the items of interest. We illustrate the framework with three exemplar studies, i.e., a case study, an interview study and a survey. There are limitations to the framework and opportunities for further research. One direction for further research is to develop guidelines or a methodology for the application of the framework in the design and conduct of field studies.

Appendix A. Summary of advice given in reviewed articles

Tables A.1 and A.2 summarise the relevant advice that was found in the ten articles discussed in Section 4.

Table A.1: Summary of the analysis of the initial six advisory sources.

Ref.	Summary
[16]	There should be a rationale for the selection of participants. Participants should give informed consent. Participants should be selected for diversity rather than similarity. Participants are not selected for statistical representation. With small samples, participants may be identifiable.
[18]	Additional participants may be selected through recommendations during interviews. Citing [38], they selected participants with first–hand experience. Overall, the guidelines focus on selecting sites rather than individuals.
[19]	Paper focuses on ethical concerns relating to participants, e.g., identifying places and settings make participants more easily identifiable. There is no guidance on recruiting participants.
[20]	It is necessary, and also probably requires a lot of effort, to select participants carefully. Participants should have free choice to participate, e.g., not influenced by their managers. Participants may drop–out, impacting the study. Recruitment of participants should be reported.
[21]	The authors highlight the importance of a sampling plan, including types of sampling. Furthermore, the guidelines describe the need for anonymity and confidentiality, as well as usability and willingness to participate.
[24]	The authors discuss how to recruit industry participants in general, e.g., carefully crafting a call for participation so as to avoid a "spam effect", and snowballing through word–of–mouth approaches, such as asking managers. They recognise the need to make specific participants requirements clear in the recruitment process, to ensure there is some benefit to the participants in doing the survey, and to collect data from participants to ensure appropriate sampling.

Table A.2: Summary of analysis of the additional four advisory sources.

Ref.	Summary
[25]	The authors highlight the recruitment of representative, insightful and motivated participants. The interactive nature of focus groups can enrich the information collected. Participants for focus groups should be carefully selected to mitigate threats, e.g., recruit participants of equal expertise.
[26]	The authors stress the importance of sampling from a *defined* population to be able to draw conclusions from the study. They discuss dropouts. They highlight the necessity of tracking the characteristics of the participants to be able to determine the effects of dropouts.
[27]	The authors highlight the need to make an informed decision of whom to include in the study.
[28]	The authors highlight the importance of ensuring that the sample is representative of the intended population.

Appendix B. Further illustrations of the framework

In this appendix, we illustrate the framework with two exemplar studies that complement the case study we considered in Section 6.2, an interview study and a survey study. A summary of all three exemplar studies is given in Section 6.3.

B.1. Interview study

The interview study we consider was published by Singer [30] as a conference paper in 1998. Singer's article concerns the maintenance of large scale software systems. Information was collected from participants through paired interviews, i.e., two participants participated in each interview. The interview questionnaire comprised three parts: background information, task analysis, and a tools wish–list. Due to time constraints and other factors, the third part of the questionnaire was rarely asked during the interview. Participants were recruited by managers. Singer does not report how many participants were recruited but indicates that participants were drawn from ten industrial sites.

Singer's article is published very "early", relative to most guidelines in software engineering. Thus, the article does not use any (stated) guidelines when making research–design decisions. Considering exemplars published *prior* to the publication of guidelines allows us to examine whether the components of the framework were implicitly recognised by at least some researchers before guidelines were established, and therefore whether the framework concerns recurring, more fundamental issues.

As with the Freimut et al. [29] article, Singer [30] also considers the nature of "experts", although not in the systematic way undertaken by Freimut et al. Singer writes, on the basis of the background information she collected, that, "These data paint a picture of software maintenance engineers as being both *expert* programmers and *experts* in the project in which they are working." ([30]). The article later further clarifies that the experts were both expert programmers and expert maintenance programmers. In terms of the R^3 model, this suggests participants whose expertise is Relevant to programming, to maintenance programming, and to the application domain/project.

We organise our discussion of the Singer [30] article around the main components of the framework:

- **Information:** The first part of the interview questionnaire collected contextual information about software maintenance (i.e., applications, languages, platforms and projects). This information may broadly be understood as collecting factual information. By contrast, the four, qualitative "truths" proposed by Singer were based on *beliefs* gathered through the second part of the questionnaire that concerned task analysis.

- **Quality**: There are some indications that Singer was at least implicitly thinking about some of these criteria. For example:

  - **Quantity of experience:** The Singer article reports information (min., max. and mean) on years experience, time on programming language, number of languages and time on project.
  - **Perceptual sensitivity:** for this criterion there are no clear examples in the Singer article.
  - **Situation selectivity:** for this criterion there are no clear examples in the Singer article, however Singer notes that, "The [ten corporate] environments themselves were diverse with respect to practically all defining variables."
  - **Reflexivity:** One reason that Singer designed the study to interview participants in pairs was to encourage "...them to verbalize

their thoughts because they could talk to each other about aspects of the project/product." Whilst Singer has not explicitly referred to reflection, the exchange of information between the two participants would, at least in principle, encourage reflection.

  - **Willingness:** Singer writes, "...it is possible that the managers chose their more stable employees to participate in the interviews." and that she used the paired–interview design to make the situation more comfortable for the participants. Whilst these two examples do not explicitly relate to willingness to share information, they would, at least in principle, encourage participants to share information.

- **Demographics**: The article refers to collecting "background information". This information relates to the projects and not to the participants. It appears therefore that no participant–related demographic information was collected.

- **The R^3 model**: Singer provides a range of comments that clearly map to the R^3 model. As examples:

  - **Real**: All participants had a minimum of three years experience, with a minimum of one year on the project.

  - **Relevant**: Singer observed that approx. 60% of the software maintainers' professional life was spent on maintenance projects and approx. 40% on new development. From this she speculates, "It is not clear if different skills are needed for these two endeavors [maintenance vs new development], but if so, then, on average, the interviewees were more familiar with the job of maintaining software programs than developing new ones." In terms of our framework, the participants' experience would therefore be more Relevant to maintenance than to new development.

  - **Recent**: Singer writes that all participants had to be working on a product that was at least 1.5 years old and "currently in a maintenance phase."

- **Roles:** Singer writes, in relation to the managers selecting the participants, "It was stressed to the managers that all participants should be involved in the actual maintenance of software (as opposed to leading a team, other administrative posts, etc.)." Singer was therefore selecting Performers.

B.2. Survey study

The survey study we consider was published by Vredenburg et al. [31] as a conference paper in 2002. Vredenburg et al.'s article concerns the use of methods, practices, key factors and trade-offs for user-centred design (UCD). The survey questionnaire was distributed to attendees at the CHI'2000 conference and then via email to members of the Usability Professional Association (UPA). 103 participants completed the survey questionnaire.

Like the Friemut et al. [29] article and the Singer [30] article, Vredenburg et al. [31] also consider experts, however Vredenburg et al. [31] refer to the participants as "opinion leaders". They write, "They [the participants] were likely opinion leaders in the UCD community, playing a leading role in their own organization's UCD practice."

Two particularly interesting aspects of the survey are, first, the way that the study recruited participants, and second, the results of the survey. For the first aspect, Vredenburg et al. [31] defined a target participant ("at least three years of experience with UCD, and considered UCD as their primary job.") and highlighted in the invitation–to–participants that only those who met the target profile should participate. Vredenburg et al. therefore encouraged prospective participants to self–select, or self–reject, themselves. Also, Vredenburg et al. [31] asked participants to consider a representative project. Vredenburg et al. [31] were therefore looking to recruit participants with at least Real and Relevant experience, as well as participants who were *Performers* in the situation of interest. Also, Vredenburg et al. [31] are distinguishing between the participant and the item of interest, e.g., the representative project.

For the second aspect, concerning the results, we consider two examples here. First, Vredenburg et al. observe a *lack of consensus* in the responses: the 103 participants identified a total of 191 indicators of UCD effectiveness. The lack of consensus, and therefore the amount of "disagreement", suggests that cultural norms were not influencing the responses. Second, whilst participants identified UCD practices that were considered useful, they were rarely used: "Only three of the [top 10] measures [for UCD success] were reported by more than 10% of the respondents and none of them was higher than

20%." This is curious for it suggests that *Performers* are not performing the identified UCD practices.

As with our two other exemplars, we organise our illustration of the mapping of the Vredenburg et al. [31] article around the main components of the framework:

- **Information:** The article provides information about project profiles (e.g., number of people on the team). Such information may broadly be understood as *factual*. By contrast, the information on, for example, measures of UCD effectiveness and applied measures, were based on *beliefs* gathered from participants who took part in the (representative) project/s.

- **Quality**: As recognised in Section 5.6 of our article, the Quality criteria are hard to measure etc. There are some indications that Vredenburg et al. were at least implicitly thinking about some of these criteria. For example:

  - **Quantity of experience:** The Vredenburg et al. article collected information on years of experience with UCD, percentage of work time on UCD-related activities over the past 12-months, number of projects involving UCD over the past 12 months, and level of familiarity with UCD practices.
  - **Perceptual sensitivity:** For this criterion there are no clear examples in the Vredenburg et al. article.
  - **Situation selectivity:** Participants were asked to select a representative project that used UCD, and in which they had participated, over the past 12 months.
  - **Reflexivity:** For this criterion there are no clear examples in the Vredenburg et al. article.
  - **Willingness:** For this criterion there are no clear examples in the Vredenburg et al. article although, presumably, completion of the questionnaire survey is an indicator of at least some willingness to share information.

- **Demographics**: The study collected limited information: country in which the participant worked, and on highest qualification (e.g., PhDs or Masters).

- **The R³ model**: Vredenburg et al. provide a range of comments that clearly map to the R³ model. As examples:

  - **Real**: Vredenburg et al. write, "... respondents appeared to be truly experienced practitioners because of their multiple years of experience and familiarity with UCD, and the fact that they attended the CHI conference or were members of the UPA."
  - **Relevant**: Vredenburg et al. asked the participants to choose a representative project. They observed that nearly 63% of the respondents chose an Internet/Intranet project.
  - **Recent**: As noted above, the article collected information on number of projects, and percentage of work time, involving UCD over the most recent 12 months. Vredenburg et al. found that on average (mean and mode) participants participated in five projects involving UCD.

- **Roles:** Vredenburg et al. write that participants were asked to select a representative project that used UCD, *and in which they had participated*, suggesting that participants were either Performers or Observers.

References

[1] S. Baltes, P. Ralph, Sampling in software engineering research: A critical review and guidelines, https://arxiv.org/abs/2002.07764 (2020). arXiv:preprint, arXiv:2002.07764v6 [cs.SE].

[2] A. Bouraffa, W. Maalej, Two decades of empirical research on developers' information needs: A preliminary analysis, in: Proceedings of the 42nd International Conference on Software Engineering Workshops, 2020, pp. 71–77. doi:10.1145/3387940.3391485.

[3] P. Runeson, M. Höst, A. Rainer, B. Regnell, Case Study Research in Software Engineering: Guidelines and Examples, John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2012.

[4] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering, Springer Science & Business Media, Berlin and Heidelberg, Germany, 2012.

[5] C. Wohlin, A. Rainer, Challenges and recommendations to publishing and using credible evidence in software engineering, Information and Software Technology 134 (June) (2021) 106555. doi:10.1016/j.infsof.2021.106555.

[6] D. Graziotin, F. Fagerholm, X. Wang, P. Abrahamsson, On the unhappiness of software developers, in: Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, 2017, pp. 324–333. doi:10.1145/3084226.3084242.

[7] V. Lenarduzzi, O. Dieste, D. Fucci, S. Vegas, Towards a methodology for participant selection in software engineering experiments: A vision of the future, in: Proceedings of the 15th International Symposium on Empirical Software Engineering and Measurement, 2021, pp. 1–6. doi:10.1145/3475716.3484273.

[8] H. R. Bernard, P. Killworth, D. Kronenfeld, L. Sailer, The problem of informant accuracy: The validity of retrospective data, Annual Review of Anthropology 13 (1) (1984) 495–517. doi:10.1146/annurev.an.13.100184.002431.

[9] D. Kronenfeld, J. Kronenfeld, J. Kronenfeld, Toward a science of design for successful food service, Institutions and Volume Feeding Management 70 (11) (1972) 38–44.

[10] B. Curtis, H. Krasner, N. Iscoe, A field study of the software design process for large systems, Communications of the ACM 31 (11) (1988) 1268–1287. doi:10.1145/50087.50089.

[11] D. Falessi, N. Juristo, C. Wohlin, B. Turhan, J. Münch, A. Jedlitschka, M. Oivo, Empirical software engineering experts on the use of students and professionals in experiments, Empirical Software Engineering 23 (February) (2018) 452–489. doi:10.1007/s10664-017-9523-3.

[12] D. Karlström, P. Runeson, Integrating agile software development into stage-gate managed product development, Empirical Software Engineering 11 (June) (2006) 203–225. doi:10.1007/s10664-006-6402-8.

[13] A. Rainer, T. Hall, N. Baddoo, Persuading developers to" buy into" software process improvement: a local opinion and empirical evidence, in:

Proceedings of the 2nd International Symposium on Empirical Software Engineering, 2003, pp. 326–335. doi:10.1109/ISESE.2003.1237993.

[14] P. Devanbu, T. Zimmermann, C. Bird, Belief & evidence in empirical software engineering, in: Proceedings of the 38th International Conference on Software Engineering, 2016, pp. 108–119. doi:10.1145/2884781.2884812.

[15] M. N. Marshall, The key informant technique, Family Practice 13 (1) (1996) 92–97. doi:10.1093/fampra/13.1.92.

[16] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empirical Software Engineering 14 (2) (2009) 131. doi:10.1007/s10664-008-9102-8.

[17] M. Höst, P. Runeson, Checklists for software engineering case study research, in: Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement, 2007, pp. 479–481. doi:10.1109/ESEM.2007.46.

[18] J. M. Verner, J. Sampson, V. Tosic, N. A. Bakar, B. A. Kitchenham, Guidelines for industrially-based multiple case studies in software engineering, in: Proceedings of the 3rd International Conference on Research Challenges in Information Science, 2009, pp. 313–324. doi:10.1109/RCIS.2009.5089295.

[19] P. E. Strandberg, Ethical interviews in software engineering, in: Proceedings of the 13th International Symposium on Empirical Software Engineering and Measurement, 2019, pp. 1–11. doi:10.1109/ESEM.2019.8870192.

[20] S. E. Hove, B. Anda, Experiences from conducting semi-structured interviews in empirical software engineering research, in: Proceedings of the 11th International Software Metrics Symposium, 2005, pp. 10–23. doi:10.1109/METRICS.2005.24.

[21] J. S. Molléri, K. Petersen, E. Mendes, An empirically evaluated checklist for surveys in software engineering, Information and Software Technology 119 (2020) 106240. doi:10.1016/j.infsof.2019.106240.

[22] J. S. Molléri, K. Petersen, E. Mendes, Cerse-catalog for empirical research in software engineering: A systematic mapping study, Information and Software Technology 105 (2019) 117–149. doi:10.1016/j.infsof.2018.08.008.

[23] J. S. Molléri, K. Petersen, E. Mendes, Survey guidelines in software engineering: An annotated review, in: Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement, 2016, pp. 1–6. doi:10.1145/2961111.2962619.

[24] N. Salleh, R. Hoda, M. T. Su, T. Kanij, J. Grundy, Recruitment, engagement and feedback in empirical software engineering studies in industrial contexts, Information and Software Technology 98 (2018) 161–172. doi:10.1016/j.infsof.2017.12.001.

[25] J. Kontio, J. Bragge, L. Lehtola, The focus group method as an empirical tool in software engineering, in: F. Shull, J. Singer, D. I. K. Sjøberg (Eds.), Guide to Advanced Empirical Software Engineering, Springer, London, 2008, pp. 93–116. doi:10.1007/978-1-84800-044-5_4.

[26] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, J. Rosenberg, Preliminary guidelines for empirical research in software engineering, IEEE Transactions on Software Engineering 28 (8) (2002) 721–734. doi:10.1109/TSE.2002.1027796.

[27] H. Zhang, X. Huang, X. Zhou, H. Huang, M. A. Babar, Ethnographic research in software engineering: a critical review and checklist, in: Proceedings of the 27th Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2019, pp. 659–670. doi:10.1145/3338906.3338976.

[28] P. Ralph, ACM SIGSOFT Empirical Standards Released, SIGSOFT Softw. Eng. Notes 46 (1) (2021) 19. doi:10.1145/3437479.3437483.

[29] B. Freimut, L. C. Briand, F. Vollei, Determining inspection cost-effectiveness by combining project data and expert opinion, IEEE Transactions on Software Engineering 31 (12) (2005) 1074–1092. doi:10.1109/TSE.2005.136.

[30] J. Singer, Practices of software maintenance, in: Proceedings of the 11th International Conference on Software Maintenance, 1998, pp. 139–145. doi:10.1109/ICSM.1998.738502.

[31] K. Vredenburg, J.-Y. Mao, P. W. Smith, T. Carey, A survey of user-centered design practice, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2002, pp. 471 – 478. doi:10.1145/503376.503460.

[32] D. I. K. Sjøberg, T. Dybå, B. C. Anda, J. E. Hannay, Building theories in software engineering, in: F. Shull, J. Singer, D. I. K. Sjøberg (Eds.), Guide to Advanced Empirical Software Sngineering, Springer, London, 2008, pp. 312–336. doi:10.1007/978-1-84800-044-5_12.

[33] K. Petersen, C. Wohlin, Context in industrial software engineering research, in: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement, 2009, pp. 401–404. doi:10.1109/ESEM.2009.5316010.

[34] T. Dybå, D. I. Sjøberg, D. S. Cruzes, What works for whom, where, when, and why? on the role of context in empirical software engineering, in: Proceedings of the 6th International Symposium on Empirical Software Engineering and Measurement, 2012, pp. 19–28. doi:10.1145/2372251.2372256.

[35] P. Clarke, R. V. O'Connor, The situational factors that affect the software development process: Towards a comprehensive reference framework, Information and Software Technology 54 (5) (2012) 433–447. doi:10.1016/j.infsof.2011.12.003.

[36] A. Begel, T. Zimmermann, Analyze this! 145 questions for data scientists in software engineering, in: Proceedings of the 36th International Conference on Software Engineering, 2014, pp. 12–23. doi:10.1145/2568225.2568233.

[37] J. C. Carver, A. Serebrenik, Gender in software engineering, IEEE Software 36 (6) (2019) 76–78. doi:10.1109/MS.2019.2934584.

[38] R. E. Stake, The art of case study research, Sage Publications, Thousand Oaks, California, USA, 1995.