User Studies on End-User Service Composition: a Literature Review and a Design Framework

User Studies on End-User Service Composition: a Literature Review and a Design Framework

LIPING ZHAO, School of Computer Science, University of Manchester, United Kingdom
PERICLES LOUCOPOULOS, Institute of Digital Innovation and Research, Dublin, Ireland
EVANGELIA KAVAKLI, Department of Cultural Technology and Communication, University of the Aegean, Greece
KELETSO J. LETSHOLO, Department of Computer Science, Botswana International University of Science and Technology, Botswana

Context: End-user service composition (EUSC) is a service-oriented paradigm that aims to empower end users and allow them to compose their own web applications from reusable service components. User studies have been used to evaluate EUSC tools and processes. Such an approach should benefit software development, because incorporating end users' feedback into software development should make software more useful and usable. **Problem**: There is a gap in our understanding of what constitutes a user study, and how a good user study should be designed, conducted and reported. **Goal**: This paper aims to address this gap. **Method**: The paper presents a systematic review of 47 selected user studies for EUSC. Guided by a review framework, the paper systematically and consistently assesses the focus, methodology and cohesion of each of these studies. **Results**: The paper concludes that the focus of these studies is clear, but their methodology is incomplete and inadequate, their overall cohesion is poor. The findings lead to the development of a design framework and a set of questions for the design, reporting and review of good user studies for EUSC. The detailed analysis and the insights obtained from the analysis should be applicable to the design of user studies for service-oriented systems as well and indeed for any user studies related to software artifacts.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods → User studies; Information systems → World Wide Web → Web services**

KEYWORDS

User Studies, Empirical Studies, Qualitative Studies, Review Framework, Design Guideline, End-User Service Composition, Web Services, Mapshups, Service-Oriented Computing, Systematic Review

ACM Reference format:

Liping Zhao, Pericles Loucopoulos, Evangelia Kavakli, and Keletso J. Letsholo. 2019. User Studies on End-User Service Composition: A Literature Review and A Design Framework. *ACM Trans. The Web.* X, XXXX, 35 pages.

1 INTRODUCTION

END-USER service composition (EUSC) is a paradigm of service-oriented computing (SOC) that aims to empower end users who are not professional developers to design or customize their own web applications [1], [2]. Such an idea can be very attractive to businesses, because providing low-cost, customized services that can respond quickly to changing requirements and environments is crucial to today's business operation. Involving end users in service composition can also help improve usability and usefulness of SOC technologies, because it will force software developers to focus on the interaction between the users and the technologies, and make the technologies better suited for users. Since the success or failure of software systems is determined by how they meet users' requirements [3], SOC developers would benefit from incorporating end users' feedback into their working practice and development process.

The user-centric, lightweight software development process of EUSC has made it an ideal platform for end-user development (EUD) [4], [5]. Yet, in spite of its potential, recent studies [6], [7], [8] found that even most advanced EUSC tools are still too difficult to use by end users, because they suffer from both conceptual and usability problems [6]. Whereas conceptual problems are concerned with understandability and learnability, which are related to the notion of "*easy to learn*" and "*easy to understand*", usability problems involve efficiency, effectiveness and appropriateness, which are related to "*easy to use*". In human-

computer interaction (HCI), such problems have been investigated by *user studies*, a user-centered design (UCD) approach for evaluating the usability and usefulness of design artifacts [9].

Although much research has been done in empirical software engineering to foster studies on usability issues, little is yet known about *what constitutes a good user study and how such a study should be designed, conducted and reported*. This paper aims to contribute to this knowledge through three steps: First, we derive a review framework for user study research from empirical study guidelines in software engineering (SE), HCI and social science. Second, we use this framework to assess a set of user studies on EUSC, selected through a systematic review process. Third, informed by the insights and results gained from the review of these studies, and in further consultation with empirical study guidelines in SE, HCI and social science, we amend our review framework and repurpose it into a design framework. We put forward this design framework as a guideline proposal, to be tested in practice by the SOC community in general and the EUSC community in particular. We intend to offer this design framework as a guideline for newcomers to user study research and to help them in the planning, designing and reporting of user studies.

In recent years, new computing paradigms, such as cloud computing, mobile computing, big data, and social computing, have drastically changed the scale and complexity of SOC. While SOC remains central to these paradigms, it needs to address new technical challenges brought up by these paradigms [10]. In the case of service composition, new challenges include precisely and efficiently searching for services from large-scale repositories (e.g., millions of mobile phone apps from cloud-based app stores), and composing a large number of services into a coherent service system that are not described by WSDL (Web Services Description Language) [10]. In the era of ubiquitous computing, we believe that the need for end-user involvement in SOC development is now more than ever of great importance. This paper is timely in revisiting our understanding of user studies on EUSC and rethinking the design issues of good quality user studies. In order not to be sidetracked to the technical issues that are raised by the aforementioned new computing paradigms, which themselves deserve proper discussions on their own right, our mapping study has chosen to focus only on one uniform set of user studies for EUSC – that is, user studies for evaluating the traditional, desktop-based web service composition, also known as web mashup [1], where the device used during service composition is a desktop.

The remaining paper is organized as follows: Section 2 introduces EUSC concepts and approaches. Section 3 introduces user study research methods and proposes a review framework for user studies. Section 4 describes our review process. Section 5 reports on the review results, whereas Section 6 answers the research questions based on the review results and discusses the key findings. In Section 7, we first critically reflect upon both the review results and the review framework, and we then use this reflection to help us derive a design framework, which we believe can be applied to user studies in general. To supplement this design framework, we also suggest the questions for designing, conducting and reporting future user studies on EUSC. Section 8 discusses the validity threats to our review and Section 9 concludes the paper.

2 OVERVIEW OF EUSC CONCEPTS AND APPROACHES

EUSC aims to enable end users, those who are not professional programmers, to create, compose and assemble Web applications from existing service components at a point of need [1]. End-user services are typically *situational software* [11], created for a narrow group of users with a unique set of needs and may have a short life span, such as tourist maps and flood maps [12]. Most such applications are in the form of "*service mashups*" [13] (also called "*web mashups"* or "*mashups*" [14] for short), which are compositions of available *web-delivered services* [15] (including SOAP-based web services, RESTful web services and RSS/Atom feeds). The term "mashup" suggests easy and fast integration of web services from multiple sources [16]. Due to their relative simplicity, mashup development practices represented a popular trend in EUSC [15].

An important characteristic of EUSC is that *users not only interact with the final software product, but are also involved in the development of the product itself* – this special dual role is seen as the empowerment of end users [17]. However, the development process involving end users is lightweight [1], comprising simple tasks such as finding relevant service components and then integrating them through mashup. This process is part of a more complex process, involving professional software developers to provide reusable

services and service components, e.g. the process of assisted composition [18]. Due to its user-centric, lightweight software development process, EUSC has been considered to be an ideal candidate for promoting end-user development [1].

EUSC approaches and tools are closely related to their underlying composition models, which range from low-level (code-based) to high-level (user interface-based) representations. These approaches and tools can be broadly classified into the following four categories [19], [17]:

- *Language-based composition* includes script-based composition such as IBM Sharable Code [20] and programming-by-example such as Vegemite [21].
- *Flow-based composition* is based on the "wire paradigm" [22] and "wired notations" [1]. Examples are Yahoo! Pipes [23] and FAST [24]). This type of tools allows users to create mashups based on data flows or control flows.
- *Form-based composition* includes template-based [25] and spreadsheet-based [8]. Examples are FormSys [26] and Karma [27].
- *Interface-based composition* is also called "webpage customization" [22]. A popular style of this composition approach is WYSIWYG (What You See is What You Get), in which mashup components are represented as icons [1]. Mashup tools that adopted this representational style are PEUDOM [28] and NaturalMash [29].

According to two user studies [1], [19], both form-based and interface-based paradigms offer representations closer to the mental model of end-users and are more appropriate for the end-user mashup development, whereas scripts and flow-based representations are most difficult to use by end users because they involve a high learning curve and demand high programming skills. The main challenge of developing EUSC tools is to provide non-technical users with an *easy to use* and *easy to understand* service composition approach, among other things.

Both notions of easy to use and easy to understand are concerned with the quality of EUSC tools. We have mapped them onto the two quality aspects of ISO/IEC 25010 standard [30], where easy to use corresponds to Product Quality and easy to understand to Quality in Use. Under ISO/IEC 25010 standard, the criteria to measure Product Quality include appropriateness, learnability, operability, accessibility, among other things, whereas the criteria to measure Quality in Use include satisfaction, effectiveness, efficiency, context coverage, etc.

In our review, we expect to see that these or similar quality criteria have been explicitly used to measure the quality of EUSC approaches and tools. In other words, the focus of the user study on EUSC should be on the assessment of the quality attributes of EUSC. In the following section, we introduce the user study research.

3 USER STUDY RESEARCH AND A REVIEW FRAMEWORK

3.1 User Study Definition

User studies are conducted in many fields for different purposes. For example, in libraries, user studies are used to find out how a particular user group obtains the information needed for the conduct of their work. In HCI, user studies are recognized as a user-centered design approach, used to learn about both *conceptual* and *usability* problems that end users face when interacting with a system [9]. In this paper, we offer a more general definition of a user study:

> **Definition**: A user study is a primary empirical study that involves the end users in the evaluation of some design artifact of use, with the intention to improve it.

Thus in the context of software development, the term "design artifact" can be a "software technology", "software tool", "software system", "user interface", "website", "feature", etc. In addition, for service-oriented computing, the term can be replaced by "composition interface", "composition approach", "mashup feature", etc.

The key to user study research is end-user involvement, which is perhaps the main characteristic that differentiates user studies from other types of empirical study such as case studies [31] and ethnographic studies [32].

In HCI, there is a type of similar study called "usability study" or "usability testing". While originally designed to address the usability issues, modern usability studies are more like user studies, as demonstrated by many influential publications on this topic (e.g., [33], [34], [35], [36], [37]). For this reason, this paper will not attempt to differentiate one from another.

User studies are founded in the field of qualitative research, because their methods are essentially descriptive and inferential in character, and their data are primarily qualitative (expressed as words or pictures) [38]. Qualitative studies often overlap [39] and user studies share many characteristics with case studies. For example, both types of study are *exploratory* in nature, their primary data are *qualitative*, and their design process is *flexible* and *reflexive* [31]. Both user studies and case studies also employ ethnographic methods [32], such as observations and interviews, for data collection.

Because user studies are similar to case studies, we can use the concepts and methods of case studies, which we know more, to understand user studies, which we know less. We convey this understanding in the following sections.

3.2 Study Goals and Research Questions

Most people start their research with a broad aim in mind. For example, they want to find out what programming concepts are difficult to learn by the beginners; or if the students should be taught object-oriented concepts first before learning Java programming; or how developers in industry use UML diagrams during software design.

During the study design, such a broad aim will evolve into a set of research questions, to be answered by the study.

The research questions of a study, which define what the researcher specifically wants to learn or understand by doing the study, are at the heart of the study design [39]. They are the one component that directly connects to all the other components of the design. More than any other aspect of the design, the research questions will have an influence on, and should be responsive to, every other part of the study. Framing good questions is the most important part of study design and researchers will have to spend quite a long time developing or modifying them. Good research questions must have a clear relationship with the study goals, because they will enable the researcher to achieve the goals; good questions should also be answerable, because there is no value to asking questions that cannot be answered.

Although there is no automatic, infallible way of generating research questions [31], Easterbrook et al. [40] suggest that researchers can use different kinds of questions such as exploratory questions, base-rate questions, relationship questions, and causality questions. They argue that understanding what kind of research question to ask is an important factor in choosing an appropriate research method. Maxwell differentiates research questions into two general types: *variance questions* and *process questions* [39]. Variance questions deal with difference and correlation; they often begin with "Is there," "Does," "How much," or "To what extent." For example, "Do exemplary medical school teachers differ from others in their teaching of basic science?" or "Is there a relationship between teachers' behavior and students' learning?" and attempt to measure these differences and relationships. Process questions, in contrast, focus on how and why things happen, rather than whether there is a particular difference or relationship or how much it is explained by other variables. For example, "how did these teachers help students learn?"

A research question may be related to a hypothesis, which is a supposed explanation for an aspect of the phenomenon under study. Hypotheses may alternatively be generated from the study conclusions for further research [31]. However, research questions cannot be replaced by hypotheses.

3.3 User Study Methods: Data Collection

"Qualitative methods focus primarily on the kind of evidence (what people tell you, what they do) that will enable you to understand the meaning of what is going on" [38]. The purpose of qualitative methods is therefore to gather evidence and to search for meaning from it. There are different types of evidence or data

source for qualitative methods, including documents, records, conversations, observational notes, and interview scripts. Accordingly, there are also different data collection methods. For user studies, as with case studies ([38], [41], [31]) and empirical studies in SE [42], commonly used data collection methods include:

- Observation
- Interview
- Focus group
- Questionnaire

These common methods are briefly described the following subsections. For a more comprehensive discussion of different types of data collection method, such as first degree, second degree and third degree of data collection, please refer to the paper by Lethbridge et al. [43].

3.3.1 Observation

Gillham [38] characterizes observation as "looking and listening", which has three basic elements:

- Watching what people do,
- Listening to what they say and
- Sometimes asking them clarifying questions.

Observational studies involve observing users performing work-related tasks in a natural setting. These studies can be approached in two general ways:

Participant observation. This approach requires the researcher to be present at the setting of the study, to directly observe (look and listen) what the study participants do and say. An important part of this kind of data collection is that the researcher keeps a written record of things observed.

Detached observation. This is the "fly on the wall" or "watching from outside" approach, which is very different from participant observation. In this method, the researcher may conduct observation through a video camera; alternatively, the researcher may use a video recorder to record the activity of the participants and watch the recording at a later time.

Gillham [38] pointed out that these two approaches are very different and they have the following main distinctions: 1) participant observation is mainly descriptive and interpretative (i.e., qualitative), whereas detached observation is mainly analytic and categorical (i.e., quantitative); 2) participant observation is subjective and humanistic, emphasizing on meaning and interpretation, whereas detached observation is objective, emphasizing on observed behavior; 3) participant observation is largely informal and flexible on information collection, whereas detached observation is formal, highly structured in data collection.

Observation can be carried out by using thinking-aloud protocols, which are a widely used method for the usability testing of software, interfaces, websites, and user manuals [44]. The basic principle of this method is that potential users are asked to complete a set of tasks with the artifact tested, and to constantly verbalize their thoughts while working on the tasks.

Observation can be used in these various ways: as an *exploratory technique* in a study; as an *initial method* in a study when other methods will take over; as the *main data collection technique* when the primary purpose is explanatory description; as *part of a multi-method* approach; and as a supplementary technique for other methods.

Observation is a most (if not the most) common method in qualitative studies, because it is the most direct way of obtaining data. See Gillham [38] for a comprehensive and detailed description of this method.

3.3.2 Interview, Focus Group and Questionnaire

These are a range of ways in which people can give the researcher information. They can be collectively called "interviewing" [38] or interview-based approaches [31] because they all share some characteristics of interviews. An interview-based approach may be informal, for example, an off the cuff spontaneous discussion, or more formal, such as a questionnaire. It can be structured as a group interview (i.e., focus group), a face-to-face individual interview, a telephone interview, or a questionnaire.

The face-to-face interview approach has the overwhelming strength due to its "richness" of the communication, but the downside is that it is time consuming. Questionnaire, on the other hand, can be

superficial and abstract, as it cannot have an in-depth or direct communication between the researcher and the respondents.

The focus group approach is particularly useful for getting an early orientation on the research topic – asking simple open questions and then noting the range and kind of responses. Issues of conflict or disagreement may indicate hidden complexities of the research. The pitfall of the focus groups is its group dynamics, which can be a powerful distorting force that either dominates proceedings or inhibit others [38]. Attention to group composition is therefore important. If a balanced group composition can be achieved, focus group has several advantages over individual interview [45]: For example, the researcher can benefit from rich discussions contributed by participants with diverse backgrounds, experiences and knowledge; the technique is also cost-effectiveness and time saver as the researcher can interview several people at the same time.

For further reading, Runeson et al. [31] provide a detailed description of interview-based approaches; Gillham [38] offers the advice on when or when not to use an interview and how to prepare for different interview studies.

3.3.3 Method Triangulation

A user study, like a case study, is a main method, whereby different sub-methods are used for data collection, including interviews, observations, and so on. If a user study collects different data by different methods but for the same studied phenomenon, then the study is said to adopt a *multi-method* approach, usually known as *method triangulation* [38], [41].

As discussed above, different methods have different strengths and different weaknesses. Triangulation is a way to *complement* different methods by taking a multiple perspective approach to data collection [31]. It helps increase the precision and strengthen the validity of empirical research [31]. The need for triangulation is particularly clear when a study relies primarily on qualitative data, because the convergence of the methods serves as an insurance policy for the researchers to be reasonably confident that they are getting a true picture of the phenomenon under study.

3.4 User Study Methods: Data Analysis

A basic principle of qualitative research is that data analysis should be conducted simultaneously with data collection [39], as this allows the researchers to progressively focus their interviews and observations, and to decide how to test emerging conclusions. Data collection and data analysis methods are two main components of qualitative methods.

While data collection methods tell the researcher how to collect data and what data to collect, data analysis methods state how the researcher can seek meaning from data. As the collected data may be specific or peculiar to the studied object, analysis has to be appropriate in order not to deform the study findings.

In qualitative research, data analysis methods are more diverse and complex than data collection methods. To start with, data analysis is conducted differently for quantitative and qualitative data, giving rise to qualitative analysis and quantitative analysis methods. Some of the key methods are described briefly in the following sections.

3.4.1 Qualitative Analysis Methods

Qualitative analysis methods are *interpretive* in nature, as they guide the researcher in the interpretation of qualitative data. There are at least three different approaches to qualitative data analysis that we consider to be useful for user studies (and case studies): *grounded theory, content analysis* and *narrative analysis* [46], [47]. Each of these approaches may use a different set of methods. Grounded theory uses three sets of *qualitative coding* procedures that help the analyst break down the original data, conceptualize it and re-arrange it in new ways [46]. The three coding procedures are termed *open coding*, *axial coding* and *selective coding*, which are used at three stages of the analysis process:

1. *Open coding*. This is the first part of the analytic process and primarily involves “fracturing”, i.e., to break down the data and to identify first level concepts and categories.
2. *Axial coding*. At this stage connections are made in new ways between categories and sub-categories.

3. *Selective coding*. This stage involves identifying one or two core categories to which all other sub-categories relate and building a conceptual framework from which to develop a grounded theory.

Content analysis is a widely used method of eliciting meaning from text (e.g., interview scripts). Its essence is "identifying substantive statements – statements that really say something [38]." Exploratory studies particularly lend themselves to content analysis in that it "gets the answers to the question to which it is applied" [46]. Software packages are available to assist with content analysis [46].

Finally, narrative analysis can be applied to any form of textual data to identify stories (or narratives). Priest et al. [46] advise how one should use narrative analysis: "First, the text is read several times. Interviewer questions and comments are deleted, as are words that detract from the key idea of each sentence or group of sentences. The remaining text is read for sense, and any further detracting words or phrases deleted. This procedure is repeated as often as necessary until fragments of themes (sub-plots) remain." Software packages are also available to assist with narrative analysis [46].

For further reading, we recommend a two-part overview of these methods by Priest et al. [46] and Woods et al. [47].

3.4.2 Quantitative Analysis Methods

Quantitative analysis methods are analytic in nature. They typically employ mathematical and statistical techniques to measure quantitative numbers. Runeson et al. [31] listed three commonly used quantitative analysis techniques for case studies:

- *Descriptive statistics*. These include mean values, standard deviations, histograms, and scatter plots. They are used to gain an overall understanding of the collected data before any other analysis methods are applied.
- *Development of predictive models*. This type of analysis is conducted in order to describe how a measurement from a later process activity is related to an earlier process measurement. This may involve using correlation analysis and regression analysis.
- *Hypothesis testing*. This type of analysis is conducted to determine if there is a significant effect of one or several variables (independent variables) on one or several other variables (dependent variables).

Quantitative analysis can be performed on qualitative data by using the (quantitative) *coding* method [42]. For example, quantitative data such as frequency count of the appearance of certain words or phrases and the number of participants in an observation study etc. can be collected from qualitative descriptions. Using a combination of qualitative and quantitative data analysis, sometimes called "mixed methods" [48], often provides a better understanding of the studied phenomenon [42].

For ease of reference, Table 1 summarizes the data collection and analysis methods discussed in this section.

3.5 Study Validity

One challenge facing researchers of qualitative studies is that they need to convince others that their study has presented a true picture of the phenomenon and the results are trustworthy. To do so, researchers need to show that there is nothing that gets in the way that threatens the validity of their research [38]. Whereas quantitative studies can design controls into the study prior to the conduct of the study, qualitative studies can only rule out most validity threats after the research has begun [39].

Table 1 Summary of User Study Methods Discussed in Section 3

Data Collection Methods	Data Analysis Methods
• Observation – Participant Observation – Detached Observation • Interview • Focus Group • Questionnaire	• Qualitative Analysis – Grounded Theory (Qualitative Coding) – Content Analysis – Narrative Analysis • Quantitative Analysis – Descriptive Statistics – Development of Predictive Models – Hypothesis Testing

Unfortunately, what types of validity threat should be considered in a qualitative study depends on what philosophical stance researchers take [39]. According to Easterbrook et al. [40], researchers with a positivist stance usually identify four types of validity, namely, construct validity, internal validity, external validity, and reliability; on the other hand, researchers with a constructivist stance see validity as trustworthiness of research results, which can be judged by credibility, transferability, dependability, and confirmability. Maxwell [39] considered two broad types of validity threat that were often raised in relation to qualitative studies should consider: researcher bias and reactivity; the second threat is the effect of the researcher on the setting or individuals studied.

3.6 Types of User Study

The term user study does not immediately convey *how* a study is conducted and *why* the study is conducted. To understand the how and why of a user study, we need to understand *what type* the user study is.

The type of a user study can be identified by its main data collection method. Thus, by labeling a user study as an observational study or an interview, it indicates *how* the user study is conducted.

The type of a user study can also be recognized by its *purpose*. Based on the purposes of case studies [31], we suggest that user studies are of three general purposes:

1. *Exploratory study*. The purpose of this type of study is to seek new insights into some phenomenon, and generate ideas and hypotheses for new research, e.g., to find out what programming concepts are difficult to learn by the beginners.
2. *Confirmatory study*. The purpose of this type of study is to confirm the ideas or test theories and hypotheses, e.g., to confirm if the students should be taught object-oriented concepts first before learning Java programming.
3. *Explanatory study*. This purpose of study aims to explain problems of some phenomena for future improvements, e.g., to explain how developers in industry use UML diagrams during software design.

Thus, by labeling a user study as an exploratory study, a confirmatory study or an explanatory study, it signifies *why* the user study is conducted.

Fig. 1 shows different types of user study, classified respectively by data collection method and by study purpose.

Understanding the purposes of user studies can also help determine *when* a particular type of user study could be conducted. Andersson and Runeson [49] show that case studies can be conducted iteratively, in alignment with the iterative project process, where the initial cycles were exploratory and the later cycles were confirmatory and explanatory.

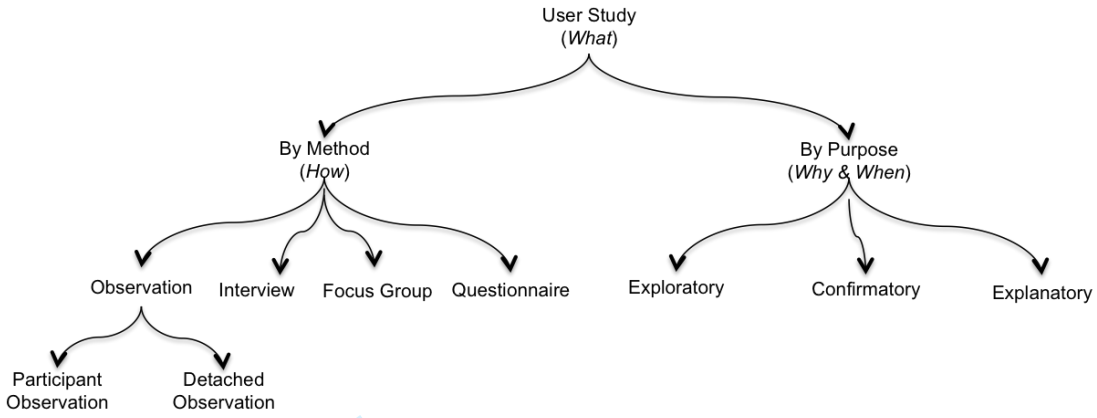


Fig. 1. Classification of user studies by data collection method versus by study purpose.

3.7 Proposal of a Review Framework for User Studies

Based on the above description, here we propose a review framework to assist the reviewer in assessing user studies. This review framework consists of eight key components of user study research, comprising Study Goals, Study Object, Study Issues, Research Questions, Data Collection Methods, Data Analysis Methods, and Study Validity. These components reflect the characteristics of a user study, as described above. In addition, these components are also common denominators of any primary empirical studies, as they can be found in many influential empirical study guidelines and design approaches in SE ([31], [41], [50], [51], [52], [53], [54]), HCI ([33], [37]) and social science ([39], [48], [55]).

To help the reviewer assess each of these eight components systematically and consistently, we have defined some prompt questions for each component. The eight user study components and their prompt questions are summarized in Table 2 and presented as follows.

- *Study Goals*. How clear is the study goal? What type of study can be recognized from the goal? It is hard to imagine that one could conduct a study without a clear sense of goals. The goals in a study serve two main functions for the study: First, they help justify the study by stating why the study is worth doing (i.e., the motivation and purpose of the study); second, they state what the researcher wants to find out in the study (i.e., the problem statement of the study). At the high level, we expect that study goals will help us, the reviewers, to understand the purpose of a study and thus enable us to label a study as an exploratory, confirmatory or explanatory study (see Fig. 1).
- *Study Object.* What type of object is being studied? What specifically about the object does the study want to focus? The study object can be anything under study [50], which provides the focus and scope of the study. In case studies, the study object is called "the case", which can be a software project or a process. In user studies, the study object can be a software tool, an user interface, a composition approach etc. [50]. Here we are particularly interested in the type of object being studied and the specific aspect of the object.
- *Study Issues.* What issues does the study want to clarify and why? The study issues of a study are the specific concerns with the study object. In user studies, study issues can be conceptual or usability issues that users have when interacting with the study object. In case studies, study issues are called "units of analysis" [31]. For example, the study of two major consecutive releases of a large legacy project can be characterized as a single embedded case study (the project) with two units of analyses (the two releases) [31].
- *Research Questions.* Have the research questions been clearly defined and explicitly answered? The importance of research questions in a qualitative study has been stated in Section 3. In this review framework, we expect that a user study should define and answer the research questions; the research questions should be related to the study goals.

Table 2. A Review Framework for Assessing User Studies

No.	Component	Types	Review Questions
1.	Study Goals	Exploratory, confirmatory, explanatory	How clear is the study goal? What type of study can be recognized from the goal?
2.	Study Object	Software technology, design artifact	What type of object is being studied? What specifically about the object does the study want to focus?
3.	Study Issues	Human-centric issues such as conceptual and usability issues	What issues does the study want to clarify and why?
4.	Research Questions	Variance or process questions, hypotheses	Have the research questions been clearly defined and explicitly answered?
5.	Study Plan	Types of user, observation tasks, interview questions, when, where	How clear is the study plan? What type of user is involved in the study? What is their background? How have they been recruited? Where is the study being conducted?
6.	Data Collection Methods	Observation (participant, detached, thinking-aloud), interview, questionnaire, focus group, or method triangulation	How does the study collect the data? What types of data are being collected? Has the study followed a standard data collection method? If so, which one?
7.	Data Analysis Methods	Qualitative analysis (e.g., qualitative coding, narrative synthesis) or mixed qualitative and quantitative analysis	How does the study analyse the data? What types of data are analysed? Has the study followed a standard data analysis method? If so, which one?
8.	Study Validity	Limitations, researcher bias, reliability, construct validity, internal, external validity	Has the study explicitly considered validity threats? What types of threat has been discussed?

- *Study Plan.* How clear is the study plan? What type of user is involved in the study? What is their background? How have they been recruited? Where is the study being conducted? This component is also called "data selection strategy" [31], in which the decisions about when and where to observe, whom to talk to, or what information sources to focus on are made [39]. Since user involvement is a defining characteristic of user studies, here we want to find out if a study has employed real users in the research.
- *Data Collection Methods.* How does the study collect the data? What types of data are being collected? Has the study followed a standard data collection method? If so, which one? According to Maxwell [39], there is no direct relationship between the research questions of the study and the data collection methods used by the study; the data collection methods are the means to answering the research questions, not a logical transformation of the latter. However, as discussed in Section 3, some data collection methods, such as observations and interviews, can provide richer data than, say, questionnaires. In addition, method triangulation is more robust than single methods, as it can obtain different data from different sources. Therefore, the way to judge the suitability of a data collection method is to see if it can give the researcher the data that are sufficient for answering the research questions and achieving the study goals.
- *Data Analysis Methods.* How does the study analyse the data? What types of data are analysed? Has the study followed a standard data analysis method? If so, which one? How the data are

analyzed has to be compatible with the research questions of the study. As stated in Section 3, primary data of user studies are qualitative. This means that the qualitative analysis should be the main type of analysis in user studies. However, as Seaman [42] suggested, complementing qualitative analysis with quantitative analysis can enrich the representation of the data. In particular, quantitative or statistical analysis can be performed on the quantitative values extracted from qualitative data (often collected from observations or interviews).

- *Study Validity.* Has the study explicitly considered validity threats? What types of threat has been discussed? In our review framework, without boiled down to any philosophical or theoretical stance of the researchers, we will be open minded by accepting a variety of criteria by which researchers may judge validity of their studies.

These eight components collectively cover two main aspects of a user study: The first four components collectively describe the *focus* of a study: What is this study about? The last four components collectively describe the *research methodology* of a study: How has the study been conducted?

This review framework will aid our review of EUSC user studies, described in the following section.

4 A SYSTEMATIC REVIEW OF USER STUDIES ON EUSC

4.1 Review Goals and Research Questions

The goal of this review is to gain a deep understanding of user studies on EUSC, to find out *what is meant by a EUSC user study*, and *how a good user study should be designed, conducted and reported*. This goal is refined into five research questions (Table 3), in accordance with our review framework.

To direct our review towards its goal, we have adopted a standard systematic review approach [56]. This approach structures the review into three phases. The first two phases, concerning the planning and conducting of the review, are described in Sections 4.2 and 4.3 respectively; the last phase, relating the reporting of the review results, is presented in Section 5. Finally, based on the review results, in Section 6 we answer our research questions.

4.2 Planning The Review

The planning phase is also called study preparation [57], which involves setting up the study design, defining appropriate research goals and questions, selecting relevant literature databases, and formulating database queries. All this amounts to the development of a review protocol [56]. The main activities in this phrase are described as follows.

Table 3. Research Questions for The Review

No.	Research Questions
RQ1	To what extent can we recognize the eight components of our review framework in the selected EUSC user studies? How cohesive are these components presented in these studies? Can we recognize any exemplars from the selected user studies?
RQ2	What type of user study is most common among the selected EUSC user studies? Which compositional approach is most studied? What issues about EUSC tools are most studied?
RQ3	What are the backgrounds of the participants of these studies? Where are they from? How have they been recruited? Where were these studies conducted?
RQ4	What is the most commonly used data collection method in these studies? What is the most commonly used data analysis method in these studies?
RQ5	What are the main characteristics of a EUSC user study?

4.2.2 Data Sources and Search Strings

The search scope for this review covered all the available primary studies published in journals and conferences in the form of user study, conducted to study some aspects of EUSE tools.

The following online databases were identified as the search sources as they were known to us to include relevant papers to this review:

- IEEE Xplore (ieeexplore.ieee.org)
- ScienceDirect (www.sciencedirect.com)
- ACM DL (dl.acm.org)
- SpringerLink (link.springer.com)

The types of papers to be searched were:

- Journals and magazines
- Conference and workshop proceedings
- Edited books (conference proceedings in Springer are published as edited books)

Based on the initial analysis of some user studies known to us, we identified the following terms for the search strings: "service composition", "mashup", "end user", and "user study".

In our search terms, we did not explicitly add specific types of user studies such as "observational study", "interview" and "questionnaire" since they were unlikely to occur independently of the term "user study".

Based on the above terms we composed a general Boolean search string *S*:

S = ("service composition" OR mashup) AND "end user" AND "user study"

We then adapted this string to the four databases based on their search engine. We found that we can apply the string as is to both IEEE Xplore and Science Direct, but we need to decompose this string into the following two substrings for ACM DL and SpringerLink:

S1 = "service composition" AND "end user" AND "user study"
S2 = mashup AND "end user" AND "user study"

The reason for this is that ACM DL's advanced search only uses "AND" (+) to query the database so we need to decompose the "OR" string into two separate strings that only contain "AND" operator. The problem with SpringerLink is different. While we appeared to be able to apply the entire string to the search engine, the search returned 39,770 results. If we limited the results to English publications, we still had 39,664 results. Clearly, we had to do something to reduce this number before we could even start study selection. We therefore decided to use two substrings to query SpringerLink and found the number of returned results manageable.

Table 4. Inclusion (I) and Exclusion (E) Criteria for Study Selection

I/E	No.	Criterion
I	1	Include the primary studies in the form of conference papers or journal articles that report user studies on some quality or usability issues of EUSC technology.
I	2	If a paper reports multiple user studies that meet the above criterion, include all the studies in the paper and treat each study as an individual study.
I	3	Where several papers report the same user study, only include the one with the most complete description.
E	4	Exclude the materials such as contents pages and editorials, white papers, commentaries, extended abstracts, and communications.
E	5	Exclude the papers that are not primary studies, including research method papers, opinion papers, and different types of review paper.
E	6	Exclude the papers that report user studies in other service composition context, such as mobile, big data, cloud computing, and IoT.

The discrepancy between different search engines used by different digital libraries has been a challenge to systematic reviews. Whilst Kitchenham et al. [56] suggested that the search strings need to be adapted to suit the specific requirements of difference databases, Brereton et al. [58] noted that current online search engines are not designed to support systematic literature reviews. We will return to this topic late in the paper, when we discuss the threats to the validity of our review in Section 8.

4.2.3 Inclusion and Exclusion Criteria

No matter how accurate search strings may be, the results returned can always contain a large number of irrelevant studies. Systematic reviews require the use of a set of explicitly defined inclusion and exclusion criteria to assess each potential primary study, to determine its relevance or otherwise and to reduce the number of relevant studies to a manageable size. We have defined three inclusion and three exclusion criteria for our review (Table 4).

4.3 Conducting the Review

Based on the defined review protocol, this phase is when the actual review is conducted. Below we describe the main steps we conducted in this phrase.

4.3.1 Database Search

This step involved using the predefined search strings to query the identified four databases to obtain a set of primary studies. The 616 search results for our review, together with the search method for each database and the filter applied, are given in Table 5. The search results were imported into an EndNote library (a bibliography management system) for study selection and review.

4.3.2 Study Selection

This step entailed the selection of the relevant primary studies from the 616 search results according to our predefined inclusion and exclusion criteria (Table 4). We conducted study selection in this order:

1. Exclude the papers according to E4, E5 and E6.
2. Include the papers according to I1, I2 and I3.

In determining the relevance of a study, we adopted the standard majority voting procedure [57], where we assigned two reviewers (the first and the last authors) to select each study independently. If a reviewer considers a study relevant, 1 point is given, 0 otherwise. After all the studies are considered this way, the studies that have 2 points are selected; the studies that receive 0 point are deselected. The studies that have 1 point are then entered into the second round of review by the third and fourth reviewers (the second and third authors) independently. At the end of the second round, the studies that have at least 2 points are selected and the studies that receive 1 point are deselected. We found that two rounds of section are sufficient.

Each reviewer followed this process to establish the relevance of a study: read the title and abstract of the study to determine if it is relevant or not; if yes, assign 1 point to the study; if not, read the full text of the study to determine its relevance; if yes, assign 1 point to the study; if not, assign 0 point to the study. We exercised caution when considering the relevance of a study purely based on its title and abstract, as the title and abstract may not tell the full story of the study.

To identify duplicate studies, we ordered the studies by author. When the same authors were found in multiple studies, each of these studies were checked against the inclusion and exclusion criteria. At the end of the third step, 45 papers were selected for the review. Among these papers two of them each reports two user studies for EUSC, thus the total number of selected studies is 47. The selected 45 papers are listed in Appendix A and their distribution over the publication channels is given in Appendix B.

Table 5. Data Sources, Search Methods and Results (Searches conducted from January 2018 to March 2018, and updated in October 2018)

Database	Search Method	Filter	Years Covered	Results
IEEE Xplore	Command search: ("service composition" OR mashup) AND "end user" AND "user study"	Metadata only (searched fields: Title, abstract, keywords)	2005 – 2018	67
ScienceDirect	Advanced search: ("service composition" OR mashup) AND "end user" AND "user study"	Computer Science journals, Publication years, web service, web application (searched fields: Title, abstract, keywords)	2005 – 2018	230
ACM DL	Advance search 1: (+"service composition" +"user study" +"end user")	Publication years (searched fields: Title, abstract, keywords)	2005 – 2018	42
	Advance search 2: (+mashup +"user study" +"end user")	Publication years (searched fields: Title, abstract, keywords)	2005 – 2018	129
SpringerLink	Command search 1: "service composition" AND "user study" AND "end user"	English (searched fields: Title, abstract, keywords)	2007 - 2018	56
	Command search 2: mashup AND "user study" AND "end user"	English (searched fields: Title, abstract, keywords)	2008 - 2018	92
Total				616

4.3.3 Data Extraction

The purpose of data extraction (i.e., data collection) is to obtain the required data that can contribute to answering the research questions of the review. We performed the following tasks for this step:

1. Read each selected study in detail to obtain an overall picture of the study.
2. Extract the descriptions from each study based on the eight components of our review framework.
3. Pull out the publication details of each study from the EndNote library.

As the data in literature review are qualitative (i.e., descriptions and diagrams), we adopted the open coding procedure (see Section 3) for the second task, where we broke down the descriptions of each study into different categories to identify their correspondence to the eight review components. We also used content analysis to identify substantive statements in each study.

We performed these tasks independently and crosschecked each other's data to ensure the interpretation of the data in relation to the eight components is correct. After we completed data extraction for all 47 studies, we merged our individual datasets into one master dataset.

4.3.4 Data Synthesis and Analysis

Data synthesis was performed as follows:

1. The data extracted from each study were sorted and aggregated into eight categories, corresponding to the eight review components.
2. Within each category, the data were sorted into different subcategories (types). For example, the data in the Study Goals category were classified into one of the Exploratory, Confirmatory and

Explanatory subcategories; the data in the Data Collection Methods category were divided into the Observation, Interview, Focus Group, and Questionnaire subcategories.

These categories and subcategories are tabulated in Appendix C.

For each category and each subcategory of data, we conducted both quantitative and qualitative analysis. The quantitative analysis produced the "headcount" for each subcategory (e.g., the number of Exploratory studies, the number of the studies that employ the Observation method, etc.). We used quantitative coding to generate the quantitative values for our review. The results of our qualitative and quantitative analysis are presented in the following section.

5 REVIEW RESULTS

In this section, we present the review results according to the eight components of our review framework.

5.1 Study Goals

Review questions: How clear is the study goal? What type of study can be recognized from the goal?

Of 47 selected studies, 39 (83%) have clearly stated their high-level goals so that we could easily recognize their study types and for the eight remaining studies, we can infer their goals from their description. We found 27 (57%) exploratory studies, 11 (23%) confirmatory studies and 9 (19%) explanatory studies (Fig. 2).

We noted that exploratory studies were concerned with the general and high-level issues, such as:

- To find out what the tool/approach can do (5)
- To seek new insights into the tool/approach (5)
- To discover problems in the tool/approach (3)
- To identify usability issues of the tool/approach (10)
- To examine the effectiveness and usefulness of the tool/approach (4)

The predominance of the exploratory studies suggests that user studies may be used primarily for exploratory purposes, a characteristic similar to case studies [31]. This may be related to the pragmatic nature of empirical software engineering research, where the practical implications of a certain practice are more relevant than the questions on abstract philosophical principles. Exploratory studies are a direct way to find out problems in practice.

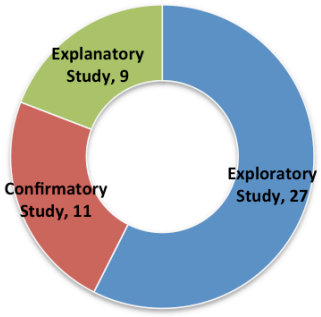


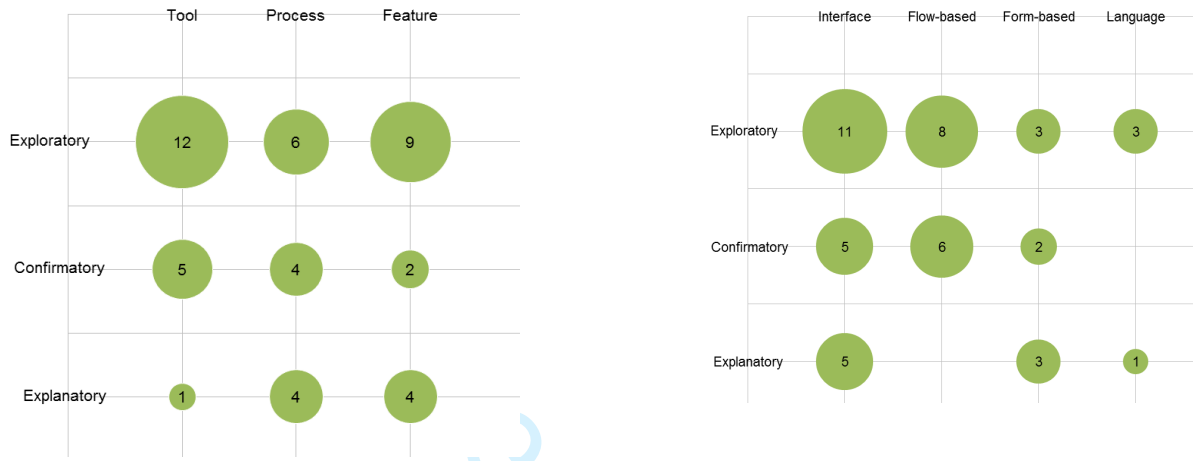
Fig. 2. The 47 EUSC user studies are classified into 27 exploratory studies, 11 confirmatory studies and 9 explanatory studies.

5.2 Study Object

Review questions: What type of object is being studied? What specifically about the object does the study want to focus?

There are three types of object being studied: EUSC Tool, Process and Feature. Fig. 3 (a) shows the number of the studies that investigated each type of the object and the distribution of the 47 studies with respect to their object types across the three study types. Fig. 3 (a) also shows that the most studied objects are EUSC tools, followed by EUSC features and processes.

We noted that a total of 16 different tools were investigated; S8-a and S8-b were the only two studies that have evaluated the same tool. This suggests that the recurrence of these tools in the 47 studies was very low. Indeed only Yahoo!Pipes occurred in five studies (S8-a, S8-b, S31, S34, S30) and the remaining tools just occurred in one or two studies at most. We also noticed that most tools described were either research prototypes or extensions of industrial tools, such as Yahoo!Pipes and Microsoft Popfly. This suggests that EUSC tools were still at an early development stage.



(a) Distribution of types of study objects across (b) Distribution of compositional approaches

Fig. 3. Distribution of different object types and compositional approaches.

Specifically for these object types, the selected studies have focused on compositional approaches. Fig. 3 (b) shows that the interface-based approach was the most studied approach, whereas the language-based approach was the least studied. Interestingly, while the form-based approach was considered to be closer to the mental model of end users [1], [22], it was less studied than the flow-based composition approach. This might be due to the popularity of some industrial EUSC tools such as Yahoo!Pipes and Microsoft Popfly, both of which are flow-based.

5.3 Study Issues

Review questions: What issues does the study want to clarify and why?

Different studies have focused on a different set of issues, but by following ISO/IEC 25010 standard [30], we have normalized the main issues in these studies into the following categories:

- Operability: To what extent can the users operate the EUSC tool to perform their tasks?
- Appropriateness: How appropriate is the tool for the users?
- Efficiency: How efficient is the tool?
- Learnability: How easy is it for end users to learn about the functionality of the tool?
- Effectiveness: How effective and accurate is the tool?
- Context Coverage: To what extent can the tool be applied consistently in many different situations?
- Understandability: To what extent can the users understand the functionality and suitability of the tool for the task at hand?

The distribution of these issues in the 47 studies is given in Table 6 and some observations are discussed below.

First, we found that the majority of these issues are related to the usability of the EUSC tools, whereas only two issues (learnability and understandability) are concerned with the conceptual problems of the tools.

Second, operability is the most studied usability issue, which were found in all 27 exploratory studies; on the other hand, understandability is the least studied conceptual issue, which were found only in the explanatory studies.

There is a strong relationship between the studies that investigate the usability issues and the exploratory studies. A possible explanation is that most exploratory studies aim to gain an understanding of the usability or usefulness of the EUSC tools, rather than other more complex, conceptual issues.

5.4 Research Questions

Review questions: Have the research questions been clearly defined and explicitly answered?

We can only find the research questions in eight studies (17%). Yet, only six of them (13%) have explicitly answered the questions.

For the rest of the studies, eight studies postulated research hypotheses at the beginning of the research and set out to test the hypotheses. As stated earlier, in qualitative studies, hypotheses are developed as tentative answers to the research questions [31], [39]. Although they are related to the research questions, they cannot be used as substitutes to the research questions.

Clearly the lack of research questions in the selected EUSC user studies is a major weakness in these studies.

5.5 Study Plan

Review questions: How clear is the study plan? How clear is the study plan? What type of user is involved in the study? What is their background? How have they been recruited? Where is the study being conducted?

Table 6. Top Study Issues Identified and Their Occurrences in the Number of Studies (out of 47)

Study Issue	No. Studies	Study Issue	No. Studies
Operability	38	Effectiveness	10
Appropriateness	23	Context coverage	5
Efficiency	21	Understandability	5
Learnability	18		

These studies have described their study plan in various details. Easily noticeable is that most studies (44/47) were conducted in a university lab and there were none conducted in an industrial setting. Correlating to this is that the main participants of these studies were made of the university students. The proportion of different types of user is summarized as follows:

- 20 studies (43%) exclusively used students;
- 5 studies (11%) exclusively used professionals;
- 10 studies (21%) used a mix of students and professionals;
- 12 studies (25%) provided no information about the types of user in their studies;
- Finally, only 4 studies (about 9%) used real users.

Most studies judged the technical backgrounds of the users according to their programming skills. 17 studies (36%) employed the users without programming skills; 4 studies (9%) employed the programmers as their users; and 22 studies (46%) used both programmers and non-programmers as their users. There were a small number of studies that reported the backgrounds of the users in terms of gender, age, education, skills, experience, and relevant training of the subject.

The users in these studies were recruited by different methods. The most common method was volunteering. Out of 47 studies, 27 (57%) used volunteers (mostly students). In eight studies (17%), money was used as an incentive to motivate people to participate, and in one study, it was mandatory for students to take part, as the study was introduced as part of the course. Lastly, in eight studies (17%), no information was provided on the recruitment method used.

Table 7 summarizes the statistics of the participants in the selected studies.

5.6 Data Collection Methods

Review questions: How does the study collect the data? What types of data are being collected? Has the study followed a standard data collection method? If so, which one?

All the studies have used a standard data collection method. The types of data collection method and the number of the studies that use them are:

- Observation-based triangulation (27 studies)
- Observation (6 studies)
- Questionnaire (5 studies)
- Interview-based triangulation (6 studies)
- Focus group (3 studies)

Observation-based triangulation was the most frequently applied approach, in which observation was used as the main data collection method, followed by an interview or questionnaire. Observation-based method triangulation was the most frequently applied data collection method. We noted that this type of triangulation has the form of *Observation with X (where X = Thinking-Aloud OR Interview OR Questionnaire)*.

The data collection procedure is generally described in detail in all the selected studies. For observational studies conducted in the lab, users were given different EUSC related tasks to perform, through which observations were made and data were collected. We have organized these user tasks into the following

Table 7 Statistics of Study Subjects in the 47 EUSC User Studies

Population	No. Studies	Background	No. Studies	Recruitment	No. Studies
Student	20	Programmer	4	Volunteer	27
Professional	5	Non-programmer	17	Paid	7
Mixed	10	Mixed	22	Compulsory	1
Unknown	12	Unclear	4	Unclear	12
Total	47	Total	47	Total	47

categories:

- *Search* – load, browse, find, retrieve, filter
- *Create* – build, make
- *Compose* – select, add, connect, merge, compile, aggregate
- *Modify* – customize, edit, update

The occurrences of these task types in different study categories are shown in Fig. 4. Note that one single EUSC user study may ask users to perform one or more types of task; conversely, not every task type is performed in every study. Thus the total occurrences of the task types in each study category can be more or less than the actual number of the studies in each category.

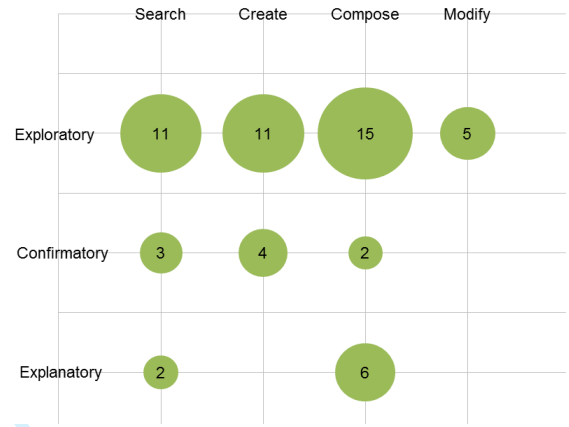


Fig. 4. Distribution of different task types across different study types. The most performed task was Compose.

Fig. 4 shows that the most performed task type was Compose, occurring in 23 (66%) of the 35 studies, and the least performed was Modify, occurring in five (14%) of the studies. The Search and Creation tasks respectively occurred in 15 (43%) and 16 (46%) of the studies. This distribution is consistent with the purpose of EUSC, which is for service composition. However, none of the studies described why the specific tasks were chosen.

Another type of triangulation is interview-based, with three forms: *Interview followed by Focus Group*, used in two studies, and *Focus Group followed by Questionnaire* and *Interview followed by Interview,* found in one study each. In total, method triangulation (both observation-based and interview-based) was used in 33 studies (70%).

By contrast, single methods were only used in 13 studies, with the observation being found in six studies, the questionnaire in five and the focus group in two. Fig. 5 shows the distribution of the data collection methods used in the 47 studies across three study types.

We believe that the high level of method triangulation in these studies is the strength of these studies. As described in Section 3, method triangulation allows for different data to be collected and can improve the validity of empirical studies. We noticed that in observation-based studies, interviews were conducted as a kind of "warm-up" session, before observations were carried out.

Overall, the observational method was the dominant data collection method, as it was used as a single method in six studies and as the main method in method triangulation used in 25 studies. This finding suggests that the current EUSC user studies were predominantly observational studies.

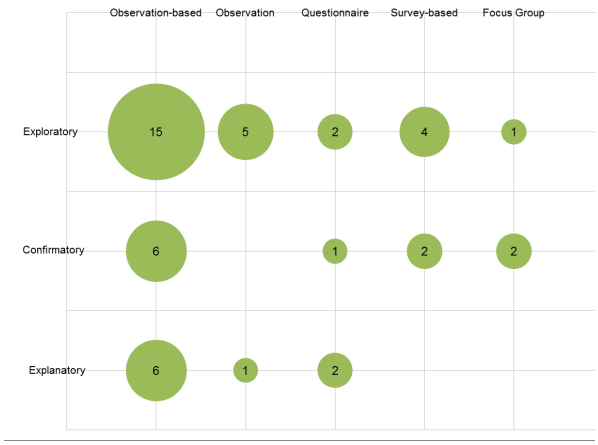


Fig. 5. Distribution of data collection methods in the 47 EUSC user studies across the three study types. Observation-based methods were most frequently used.

5.7 Data Analysis Methods

Review questions: How does the study analyse the data? What types of data are analysed? Has the study followed a standard data analysis method? If so, which one?

In contrast to data collection methods, descriptions of data analysis methods were implicit in the selected studies. By examining how these studies analyzed and presented data, we inferred the following data analysis methods from these studies:

- Descriptive statistics (18 studies)
- Narrative analysis (13 studies)
- Mixed analysis (11 studies)
- Hypothesis testing (5 studies)

Based on our inference, descriptive statistics (quantitative analysis) was the most used analysis method, used in 18 studies (38%). This seems to confirm the statement made by Runeson et al. [31] that descriptive statistics is often a natural step before any other analysis methods are applied.

Overall, quantitative analysis (i.e., descriptive statistics and hypothesis testing) is the most used analysis type, found in 23 out of 47 studies. The remaining 24 studies were split more or less evenly, with 13 using qualitative analysis (i.e., narrative analysis) and 11 using a *mixed analysis* method. Mixed analysis methods are of two types: descriptive statistics complemented with narrative analysis (10 studies) and hypothesis testing combined with narrative analysis (one study). Fig. 6 shows the distribution of different data analysis methods in the 47 studies across three study types.

We noted the following relationships between these analysis methods and the types of data they analyze in the studies:

- Descriptive statistics and hypothesis testing were always applied to *quantitative data*, regardless of their usage as a single method or as part of a mixed method.
- Narrative analysis, when used as part of a mixed method, was applied to both *qualitative and quantitative data*. The way this method processed quantitative data was through translation and interpretation, and the way it handled qualitative data was by narrative summary.
- However, when used by itself, narrative analysis was only applied to qualitative data, to provide a narrative summary. Furthermore, when used alone, narrative analysis tended to be employed in a superficial way, to provide a brief summary of the results, as demonstrated in studies S35, S36, S37, S38, and S39.

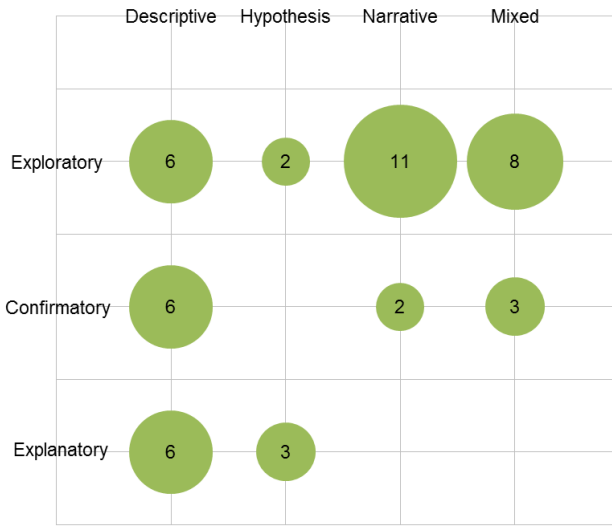


Fig. 6 Distribution of data analysis methods in the 47 EUSC user studies across the three study types. Descriptive statistics was used most. Quantitative methods (i.e., descriptive statistics and hypothesis testing) were more popular than qualitative methods (narrative synthesis).

An interesting observation is that there is an asymmetric relation (or anti-symmetry) between method triangulation and mixed analysis methods, in that while method triangulation was leading in data collection, mixed analysis was trailing behind. This might be regarded as a weakness in the current EUSC studies.

5.8 Study Validity

Review questions: Has the study explicitly considered validity threats? What types of threat has been discussed?

Only five studies (11%) have explicitly considered validity threats. These studies are S8-a, S8-b, S16, S22, and S45. This is the weakest component in the selected studies.

Four of these five studies discussed construct, internal and external validity threats, which are in line with a positivist stance. One of them, S16, considered internal, external and statistical validity, where statistical validity is similar to reliability validity, so we can say that this study also took a positivist stance. Easterbrook et al. found that survey research and case studies in SE are frequently conducted with a positivist stance, as influenced by controlled experiments [40].

6 RESEARCH FINDINGS

In this section, we use the review results to answer our research questions.

RQ1: To what extent can we recognize the eight components of our review framework in the selected EUSC user studies? How cohesive are these components presented in these studies? Can we recognize any exemplars from the selected user studies?

We can only find all eight components in three studies (S8-a, S8-b and S45). The remaining studies either miss the Research Questions (39/47) or Study Validity (42/47) or both of these components (39/47). This means that nearly 94% of the studies (44/47) have explicitly described the remaining six components.

To assess the cohesion of these components in the selected studies, we focus on the following inter-component relationships in these studies (similar to chain of evidence [41]):

1. Is there a clear link between these questions and the study issues?
   This link is clear only to the eight studies that have posed the research questions.
2. Is the data collection method used in the study appropriate to the study goal?
   Yes, there is a clear link between the study goal and data collection method in all the studies.
3. Is the data analysis method used in the study appropriate to the collected data?
   Not clear, as none of the selected studies gives an explicit justification on why they have used a certain data analysis method.
4. Is there a clear link from the collected data to the study conclusions?
   No, this connection has not been explicitly established in the studies.
5. Is there a clear link between the study tasks and the study issues?
   Partially. Most studies did not provide a clear definition of their study issues. Consequently, whether the tasks performed are appropriate is not clear.

Our conclusion is therefore that the 47 EUSC user studies are not very cohesive with respect to the eight components of our review framework.

Finally, we regard the three studies (S8-a, S8-b and S45) as exemplary user studies on EUSC (Table 8), not only because they contain all eight components, but also because these components are cohesive in these studies. Interestingly, all three studies were reported by the same set of authors. A summary of these three studies is provided in Appendix D.

RQ2: What type of user study is most common among the selected EUSC user studies? Which compositional approach is most studied? What issues about EUSC tools are most studied?

Based on our review results (Section 5.1), Exploratory Study is the most common type of study, found in more than half of the selected studies (57%), whereas Explanatory Study is the least common type, found in less than one fifth of the studies (19%).

According to Section 5.2, the interface-based approach is most studied, found in 21 studies. Of these studies, 11 are exploratory.

As discussed in Section 5.3, the majority of the issues are related to the usability of the EUSC tools, whereas only two issues (learnability and understandability) are concerned with the conceptual problems of

Table 8. Exemplary EUSC User Studies

S8-a & S8-b	S. K. Kuttal, A. Sarma, and G. Rothermel, "On the benefits of providing versioning support for end-users: An empirical study," ACM Transactions on Computer-Human Interaction, vol. 21, pp. 1-43, 2014.
S45	S. K. Kuttal, A. Sarma, G. Rothermelc, and Z. Wang, "What happened to my application? Helping end users comprehend evolution through variation management," Information and Software Technology, vol. 103, pp. 55-74, 2018

the tools. In addition, operability is the most studied usability issue, which were found in all 27 exploratory studies; on the other hand, understandability is the least studied conceptual issue, which were found only in the explanatory studies. This could be an indication that most exploratory studies were concerned with the issues closely related to usability, rather than conceptual.

RQ3: What are the backgrounds of the participants of these studies? Where are they from? How have they been recruited? Where were these studies conducted?

Based on Section 5.5, four studies (less than 10%) are found to have exclusively used programmers; 17 studies (about 36%) have exclusively used non-programmers; 21 studies (about 45%) have used both types of participants. The remaining four studies have not declared the backgrounds of their participants.

The participants of these studies are drawn from the student and professional communities, with the students as the main source of participants, who have been exclusively used in 19 studies. Professionals have been exclusively used in five studies, whereas a mix of both students and professionals are used in 10 studies. There are 12 studies that have not explicitly specified where their participants are from.

Most studies (more than half) have recruited volunteers in their studies; 7 studies have paid their participants and one study was carried out as part of a course so their participants are mandatory. There are 12 studies that have not explicitly stated how they have recruited their participants.

The above answer shows that more than a quarter of the studies have not adequately described the study plan.

RQ4: What is the most commonly used data collection method in these studies? What is the most commonly used data analysis method in these studies?

Our review found that 33 studies (70%) have either used an observation or an observation-based method for data collection. For data analysis, about a half of the studies have used qualitative (narrative analysis) or qualitative-based analysis (mixed method) and another half have used quantitative analysis (descriptive statistics and hypothesis testing).

This asymmetric relationship between data collection and data analysis methods in the selected studies is concerning, because all the selected studies have used a qualitative method for data collection, but when it comes for data analysis, a half of these studies have solely used a quantitative method. We consider the lack of using qualitative analysis in the selected studies to be another major limitation in the selected studies. Seaman [42] argued that nearly any software engineering issue is best investigated using a combination of qualitative and quantitative methods. Since user studies are qualitative, qualitative analysis methods should be the primary analysis methods in these studies.

RQ5: What are the main characteristics of a EUSC user study?

Based on our answers to RQ2 – RQ4, we suggest that a typical EUSC user study should have the following characteristics:

- Primary goal: Exploratory
- Study object: EUSC tool
- Primary study issue: Usability
- Primary participant: Non-programmer (i.e., ordinary user)
- Primary data collection method: Observation
- Primary data: Qualitative
- Primary analysis method: Qualitative

7 REFLECTION AND PROPOSAL OF A DESIGN FRAMEWORK FOR EUSC USER STUDIES

In this section, we first reflect on our review and then propose a guideline for designing, conducting and reporting future EUSC user studies.

7.1 Reflection

7.1.1 Reflecting on Review Results

Our review results have shown some major limitations with the selected EUSC user studies. First, more than 80% of the selected studies miss the Research Questions or the Study Validity component in their report. We do not believe that this omission is intentional; rather, we believe that this indicates a lack of a clear guideline on what should be included in a user study report.

Second, the description of the Study Plan component is not complete, as more than a quarter of the studies have not adequately described this component.

Third, the descriptions of data collection and analysis methods in these studies are not detailed enough, particularly on the 'how' and 'why' aspects. The main reason for this might be down to page limit imposed on the reports of these studies, especially if the study has been published as a conference paper. It could also be that the researchers did not know what should or should not be reported and how much it should be reported. In the latter case, a clear guideline would be helpful. Finally, a half of the studies have not used qualitative data analysis methods.

The question is: What should be included in the guideline and in what form?

7.1.2 Reflecting on our Review Framework

Our review framework was formulated based on many influential guidelines for qualitative studies. The framework is component-based, which emphasizes the separation of concerns in a study. As Yin [59] states, "Every type of empirical research has an implicit, if not explicit, research design." Our intention for this review framework was therefore to make the design of user study explicit, to get it out in the open, where its strengths, limitations, and implications can be clearly understood [39]. So our framework has served its purpose in this regard, by helping us review each of the 47 selected studies in a systematic manner.

The question is: Can we reposition this review framework so that it can be used as a design framework for user study design?

The answer is we can, because the report of a study should reflect the study itself and the review framework should also at least partially reflect the actual structure of the study itself.

But, as is, through our use of this framework, we have identified two major weaknesses in it: On the one hand, the framework is not concise enough and two of its eight components – Study Object and Study Issues – are overlapping. We believe these two components should be combined into one, which can be called "Study Problem".

On the other hand, the framework misses one important component, that of the theoretical frame of reference that guides and informs the study [39], [31].

Maxwell calls such a frame of reference "conceptual framework", which is the system of concepts, assumptions, expectations, beliefs, and theories that researchers draw on for understanding the people or issues they are studying [39]. It gives a tentative theory of what you think is going on with the phenomena you are studying and why. However, using theories to underpin empirical research is not yet well established in SE and Runeson et al. [31] suggested that related work can be used as a temporary alterative to the theory. This seems to be an accepted practice in qualitative research, as Maxwell [39] states that prior research findings, preliminary studies, and personal experiences can all be used to inform the study.

Fernandez and Passoth [60] posit that empirical software engineering is shifting from a single discipline to an inter-discipline "where social, cultural, and human-centric issues shape the configurations of questions, research methods, and teams". They point out, "When transferring approaches, concepts, and methods from other disciplines, we not only adopt their application, but also the underlying theories." They call this "symmetrical collaboration".

As user study research is interdisciplinary in nature, drawing on approaches, concepts, and methods from SE, HCI and social science, researchers of user studies should explain at least what literature or related work has influenced their study.

In addition, we found that the Study Plan component is inadequate, as it has not considered how the collected data will be recorded, stored and managed. We believe such a data management plan is needed to ensure data quality and transparency. Furthermore, we suggest that the Study Plan explicitly includes a study schedule, which considers the time, duration, place, environment, and timelines of the study. We believe that information about the time, duration and place of the study is important to a user study (and also any empirical study), especially when the study is intended for future replication [61]. However, this information is often omitted in empirical studies in both SE and HCI. We also propose to make the user recruitment plan as an explicit element in the Study Plan.

7.2 Repositioning the Review Framework for User Study Design

To reposition our review framework as a design framework for user study research, our above discussion suggests that the review framework should undergo the following changes:

- To combine the Study Object and Study Issues components into one component called "Study Problem"
- To introduce a new component called "Theory"
- To enhance the Study Plan component with a user recruitment plan, a study schedule and a data management plan.

The resulting design framework contains eight design components (Table 9), with the first four covering the Study Focus facet and the last four concerned with the Research Methodology. Similar to the review framework, this design framework also provides a set of questions or design decisions for each component. These design decisions act as a series of interrogatives, concerning *"what", "why", "how", "who", "when",* and *"where"*. These questions refer to the user study researcher as "you", to make it clear that "you", the researcher, are responsible for answering these questions.

7.3 Using the Design Framework for Reporting and Reviewing User Studies

From the reader or reviewer's perspective, a study is judged solely by its report. Runeson and Höst [41] state: "An empirical study cannot be distinguished from its reporting. The report communicates the findings of the study, but is also the main source of information for judging the quality of the study." Researchers should therefore always have readers in their mind when presenting their studies.

Furthermore, as we suggested above, the structure of a study should be reflected in the structure of its report, the idea behind our adaptation of the review framework to a design framework. This means that we can use the same framework for three purposes: for designing, reporting and reviewing user studies. While we can use the same set of components for all these purposes, we need to use two sets of questions, one for design and one for reporting and reviewing. Table 10 lists the questions for these latter two purposes.

Table 9. A Design Framework for User Study Research

No.	Component	Types	Design Questions
1.	Study Goals	Exploratory, confirmatory or explanatory	Why do you want to conduct this study? Why is the study worth doing? What do you want to achieve? What type of study (exploratory, confirmatory or explanatory) do you want to conduct?
2.	Theory	Underlying theories, philosophical stance, related work, prior research findings, preliminary studies, personal experiences	What theories, beliefs, and related work will guide or inform your research? What literature, preliminary studies, and personal experiences will you draw on for understanding the phenomenon you are studying?
3.	Study Problem	Human-centric issues such as conceptual and usability issues	What problem do you want to address? Has the problem been addressed before? Why is the problem worth addressing?
4.	Research Questions		What do you want to discover? What questions do you want to answer? Are these questions related to your research problem?
5.	Study Plan	User recruitment plan, study schedule and data management plan	How are you going to recruit the participants (users) for your study? What type of user will be recruited and how many? What is your study schedule (where, when and for how long)? How are you going to manage your study data?
6.	Data Collection Methods	Observation (participant, detached, thinking-aloud), interview, questionnaire, focus group, or method triangulation	Where is your data source? How do you collect data (observation, interview or triangulation etc.)? Will the data collected be sufficient to answer your research questions?
7.	Data Analysis Methods	Qualitative analysis (e.g., qualitative coding, narrative synthesis) or mixed qualitative and quantitative analysis	How do you analyze the data (qualitative or both qualitative and quantitative)? How are you going to present the analysis results (tables, graphs etc.)? How are you going to answer the research questions?
8.	Study Validity	Study results, conclusions, study limitations, researcher bias, reliability, construct validity, internal, external validity	Do your study results adequately answer your research questions? Why should we believe the results? How might they go wrong? What are the potential validity threats to them? How will you deal with the threats?

8 STUDY VALIDITY AND LIMITATIONS

In this section, we reflect on the potential threats to the validity of our review and the limitations of the review.

1. Systematic Literature Review vs. Systematic Mapping Study. As a novice, we found that the boundaries between these two types of study are not clearly defined. At first we considered our review to be a mapping study, because it involved building a classification schema for types of user study and types of study method etc. However, our reviewers have pointed out that what we have done was beyond the mapping study because we also had to review each selected study in detail in order to understand its various components. One of the reviewers recommended us the article by Kuhrmann et al. [57], which provides a clear distinguish between these two types of study. The article also suggests that both types of study can be conducted in combination. Based on this understanding, our review is a mix of a systematic literature review (SLR) and a mapping study. However, in comparison with a pure mapping study, our mapping study has not provided a

Table 10. Using the Design Framework for Reporting and Reviewing User Study Research

No.	Component	Types	Questions for Reporting and Review
1.	Study Goals	Exploratory, confirmatory or explanatory	What is the purpose and type of the study?
2.	Theory	Underlying theories, philosophical stance, related work, prior research findings, preliminary studies, personal experiences	Is the study based on a theory or related work?
3.	Study Problem	Human-centric issues such as conceptual and usability issues	What problem does the study want to investigate? Why is the problem important?
4.	Research Questions		Has the study clearly defined the research questions and explicitly answered them? Are the questions appropriate and linked to the research problem?
5.	Study Plan	User recruitment plan, study schedule and data management plan	How clear is the study plan? What type of user is involved in the study? What is their background? How have they been recruited? Where is the study being conducted and when? How are the data going to be managed?
6.	Data Collection Methods	Observation (participant, detached, thinking-aloud), interview, questionnaire, focus group, or method triangulation	How does the study collect the data? What types of data are being collected? Has the study followed a standard data collection method? If so, which one?
7.	Data Analysis Methods	Qualitative analysis (e.g., qualitative coding & narrative synthesis) or mixed qualitative and quantitative analysis	How does the study analyse the data? What types of data are analysed? Has the study followed a standard data analysis method? If so, which one?
8.	Study Validity	Study results, conclusions, study limitations, researcher bias, reliability, construct validity, internal, external validity	Have the study results adequately answered the research questions? Why should we believe the results? What are the potential validity threats to them? How has the study dealt with the threats? Does the report provide a clear chain of evidence from the goals to data collection to analysis to conclusion? Is the report coherent, easy to read and well structured?

big picture of the publication space [57] for EUSC user studies; instead, the purpose of our mapping study has been to provide a structured overview of the selected studies. In our review, we have followed the guidelines for performing SLRs in software engineering [56] and the guidelines for conducting systematic mapping studies [62], [63]. We therefore do not consider our mix of a mapping study and a SLR to be a validity threat to our review.

2. Literature Search. The most difficult task in identifying relevant EUSC user studies is the construction of the search strings, because different search engines of different digital libraries work differently and have different requirements on search strings [58]. To tame the search engines that we were using, we iteratively refined our search strings through many trial searches. During the initial trial, we tested if the search strings could identify all the papers that were always known to us. The subsequent trial searches were conducted to ensure that the search strings could find all the relevant papers cited in the known papers, a kind of

"backward snowballing" search [64]. As described in Section 4.2, we also decomposed our master search string into two substrings for ACM DL and SpringerLink. We noted that other researchers also used different search strings to cater for different search engines [63]. However, with hindsight, our search strings could be refined further, to consider study context and domain [65]. Due to this limitation, we will have inadvertently and inevitably missed many relevant studies. This is a common limitation to all literature reviews (including both systematic and non-systematic reviews), because it is simply not possible to find every relevant study. Wohlin et al. [66] and Petersen et al. [63] argued that since the actual population of all relevant studies is unknown, researchers should aim at finding a representative sample the population.

3. Searched Databases. We have made an informed decision to focus only on four major digital libraries, as we knew that contain relevant studies. There are many other libraries that we have not searched, including Oxford University Press Library, Wiley, Indersciences, and IGI Global. This is clearly a limitation of our review as our search coverage is not comprehensive. However, no review can possibly cover all the libraries, due to time constraints and page limits on papers. We do not consider Google Scholar to be suitable for literature review, as it was designed to search a wide range of documents and could return hundreds or even thousands of irrelevant papers.

4. Study Selection. Once the search results were obtained, selecting the relevant EUSC studies also posed a potential threat to our study, as we were facing hundreds of potentially relevant papers. We have mitigated this threat by taking the following countermeasures: 1) carefully designing a set of comprehensive inclusion and exclusion criteria supported by a selection process and 2) scrutinizing the papers written by same authors to ensure no duplicate studies were included in our selected studies. To determine the relevance of each study, we have adopted a procedure similar to the majority voting [57] to vote the study, as described in Section 4.

5. Data Extraction. Systematic reviews in SE are qualitative research as they deal with natural language descriptions [58]. The data to be extracted are also mostly qualitative, as they are descriptive. One potential threat to data extraction is that data can be distorted by the researcher's misinterpretation. This threat can be difficult to avoid in SE literature reviews due to the lack of standard terminology and standards for reporting experiments [67]. Common counter-measures are to have two researchers to extract the data independently [67], or to have one researcher acting as data extractor and another acting as data checker [58].

We combated the data extraction threat at two levels: At the conceptual level, our review framework provides a detailed specification and terminology for each data item we were looking for. The framework also served as a map to navigate us through the unstructured text of each study.

At the procedural level, all the researchers acted as both data extractor and data checker. We divided the data extraction task vertically for the whole collection of the 47 studies and used the same data extractor for the same data item. We recorded the extracted data on Google Sheets to facilitate crosschecking and processing of data. When a data extractor could not determine the type of a study aspect, he or she would leave a comment on Google Sheets, which would then be dealt with by the fellow data extractors. Extracted data were double or triple checked by all researchers; doubts and questions were dealt with continuously and timely through weekly consensus meetings via Skype. After consensus meetings, only a small proportion of data items needed to be rechecked again. However, our detailed and rigorous data extraction was possible because we were only dealing with a relatively small set of studies.

6. Data Synthesis. The main threats to the validity at this stage are potential researcher bias and statistical errors. Our countermeasures for these threats were similar to those of data extraction. We divided the data synthesis task vertically for the whole collection of the 47 studies and used the same researcher for the same data item. We used the spreadsheet functions on Google Sheets to automatically count and aggregate the numbers.

To ensure the accuracy of data synthesis, we revisited each study to do fact checking. At the end we believe we have high confidence in the robustness of the synthesized review results.

7. Data Analysis. The main validity threat to data analysis is related to statistical errors. To avoid this threat, we conducted data analysis meticulously, and the results were checked and rectified independently by each of us. In addition, our predefined review framework provided a standard structure for results analysis and ensures that we analysed each study element consistently and systematically.

8. Study Scope. One limitation of this review is our choice of the type of EUSC user studies, that is, we only focused on user studies on desktop based EUSC technologies. Our intention was to restrict our investigation to a homogeneous group of EUSC user studies so that we could have a level playing field for selecting and reviewing relevant studies. Our future work will investigate EUSC user studies in other computing paradigms.

9 CONCLUSIONS

Motivated by our desire to provide a good understanding of what constitutes a user study for EUSC and how such a study should be designed, conducted and reported, we have embarked on the following journey:

First, from many influential guidelines for empirical studies, we have synthesized a robust review framework. While the original purpose of this framework was to allow us to systematically and consistently characterize and assess the 47 selected EUSC user studies, we believe it should be more general than it was intended, because it was built on eight fundamental components of empirical studies.

Second, based on this framework we have conducted a systematic review of the 47 relevant user studies for EUSC to assess their focus, methodology and overall cohesiveness. The assessment has revealed the fundamental problems of these studies, which include missing the research questions in more than 80% of the studies, lack of important details in describing data collection and analysis methods in more than two thirds of the studies, missing study validity discussion in nearly 90% of the studies, and lack of strong cohesion in individual studies. Although the review only focused on the EUSC user studies, the detailed analysis of which and the manner of the analysis should provide useful insights applicable to user studies in general.

We should, however, make it clear that it was not our intention to judge the real value or contribution of the selected studies; instead, our aim was to find out if and how many of these studies have explicitly described the eight components that we consider to be essential for any empirical studies.

Third, informed by these findings and reflecting on our review experience, we have repositioned the review framework for study design and consolidated its components, to provide a guideline for the design, reporting and reviewing of a good EUSC user study. Although the design framework and its associated design questions were recommended specific to EUSC user studies, as a remedy to the current state of these studies, they should be more generally applicable to user studies for interactive systems, because they are grounded on general empirical study guidelines. However, we put forward this design framework as a proposal for the SOC or wider communities to test, debate and improve.

To conclude, this paper has made, in our view, four important contributions: (i) a detailed characterization of user study research, (ii) the development of a review framework for assessing user studies on EUSC, (iii) the presentation of the first ever systematic review of EUSC user studies, and (iv) the proposal of a more general design framework for user studies.

Acknowledgements

We are really grateful for the expert comments and excellent advice we have received from our three reviewers. Their thoughtful guidance has really helped in strengthening the manuscript.

REFERENCES

[1] F. Daniel and M. Matera, *Mashups: Concepts, Models and Architectures*: Springer, 2014.

[2] L. Xuanzhe, M. Yun, H. Gang, Z. Junfeng, M. Hong, and L. Yunxin, "Data-Driven Composition for Service-Oriented Situational Web Applications," *Services Computing, IEEE Transactions on,* vol. 8, pp. 2-16, 2015.

[3] B. Nuseibeh and S. Easterbrook, "Requirements Engineering: A Roadmap," in *Future of Software Engineering*, 2000, pp. 35-46.

[4] H. Lieberman, F. Paternò, M. Klann, and V. Wulf, *End-user development: An emerging paradigm*: Springer, 2006.

[5] D. Tetteroo and P. Markopoulos, "A Review of Research Methods in End User Development," in *End-User Development*. vol. 9083, P. Díaz, V. Pipek, C. Ardito, C. Jensen, I. Aedo, and A. Boden, Eds., ed: Springer International Publishing, 2015, pp. 58-75.

[6] A. Namoun, T. Nestler, and A. De Angeli, "Conceptual and Usability Issues in the Composable Web of Software Services," in *Current Trends in Web Engineering*. vol. 6385, F. Daniel and F. Facca, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 396-407.

[7] A. Namoun, T. Nestler, and A. De Angeli, "Service Composition for Non-programmers: Prospects, Problems, and Design Recommendations," in *Web Services (ECOWS), 2010 IEEE 8th European Conference on*, 2010, pp. 123-130.

[8] Ž. Obrenović and D. Gašević, "End-user service computing: Spreadsheets as a service composition tool," *IEEE Transactions on Services Computing,* vol. 1, pp. 229-242, 2008.

[9] W. M. Newman, M. G. Lamming, and M. Lamming, *Interactive system design*: Addison-Wesley Reading, 1995.

[10] A. Bouguettaya, M. Singh, M. Huhns, Q. Z. Sheng, H. Dong, Q. Yu*, et al.*, "A service computing manifesto: the next 10 years," *Communications of the ACM,* vol. 60, pp. 64-72, 2017.

[11] S. Balasubramaniam, G. Lewis, S. Simanta, and D. B. Smith, "Situated software: concepts, motivation, technology, and the future," *Software, IEEE,* vol. 25, pp. 50-55, 2008.

[12] S. Aghaee and C. Pautasso, "Mashup development with HTML5," in *Proceedings of the 3rd and 4th International Workshop on Web APIs and Services Mashups*, 2010, p. 10.

[13] D. Benslimane, S. Dustdar, and A. Sheth, "Services mashups: The new generation of web applications," *IEEE Internet Computing,* pp. 13-15, 2008.

[14] C. Cappiello, F. Daniel, M. Matera, M. Picozzi, and M. Weiss, "Enabling end user development through mashups: requirements, abstractions and innovation toolkits," in *International Symposium on End User Development*, 2011, pp. 9-24.

[15] J. Yu, B. Benatallah, F. Casati, and F. Daniel, "Understanding mashup development," *IEEE Internet computing,* vol. 12, 2008.

[16] D. Lizcano, J. Soriano, M. Reyes, and J. J. Hierro, "EzWeb/FAST: reporting on a successful mashup-based solution for developing and deploying composite applications in the upcoming web of services," in *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, 2008, pp. 15-24.

[17] F. Daniel and M. Matera, "Mashups and End-User Development," in *Mashups*, ed: Springer Berlin Heidelberg, 2014, pp. 237-268.

[18] N. Mehandjiev, F. Lecue, U. Wajid, and A. Namoun, "Assisted Service Composition for End Users," in *Web Services (ECOWS), 2010 IEEE 8th European Conference on*, 2010, pp. 131-138.

[19] S. S. Minhas, P. Sampaio, and N. Mehandjiev, "A Framework for the Evaluation of Mashup Tools," in *Services Computing (SCC), 2012 IEEE Ninth International Conference on*, 2012, pp. 431-438.

[20] E. M. Maximilien, A. Ranabahu, and K. Gomadam, "An Online Patform for Web APIs and Service Mashups," *IEEE Internet Computing,* vol. 12, 2008.

[21] J. Lin, J. Wong, J. Nichols, A. Cypher, and T. A. Lau, "End-User Programming of Mashups with Vegemite," in *Proceedings of 14th International Conference on Intelligent User Interfaces*, 2009, pp. 97-106.

[22] S. S. Minhas, P. Sampaio, and N. Mehandjiev, "A framework for the evaluation of mashup tools," in *Proceedings of 9th IEEE International Conference on Services Computing (SCC)*, 2012, pp. 431-438.

[23] M. C. Jones and E. F. Churchill, "Conversations in developer communities: a preliminary analysis of the yahoo! pipes community," in *Proceedings of the fourth international conference on Communities and technologies*, 2009, pp. 195-204.

[24] D. Lizcano, F. Alonso, J. Soriano, and G. López, "A component- and connector-based approach for end-user composite web applications development," *Journal of Systems and Software,* vol. 94, pp. 108-128, 8// 2014.

[25] N. Mehandjiev, A. Namoune, U. Wajid, L. Macaulay, and A. Sutcliffe, "End User Service Composition: Perceptions and Requirements," in *Web Services (ECOWS), 2010 IEEE 8th European Conference on*, 2010, pp. 139-146.

[26] I. Weber, H.-Y. Paik, and B. Benatallah, "Form-Based Web Service Composition for Domain Experts," *ACM Trans. Web,* vol. 8, pp. 1-40, 2013.

[27] R. Tuchinda, C. A. Knoblock, and P. Szekely, "Building Mashups by Demonstration," *ACM Trans. Web,* vol. 5, pp. 1-45, 2011.

[28] M. Matera, M. Picozzi, M. Pini, and M. Tonazzo, "PEUDOM: A mashup platform for the end user development of common information spaces," in *Web Engineering*, ed: Springer, 2013, pp. 494-497.

[29] S. Aghaee and C. Pautasso, "End-User Development of Mashups with NaturalMash," *Journal of Visual Languages & Computing,* vol. 25, pp. 414-432, 8// 2014.

[30] ISO/IEC, "Systems and software engineering - Systems and software quality requirements and evaluation (SQuaRE) - System and software quality models," ISO/IEC FDIS 25010, 2011.

[31] P. Runeson, M. Host, A. Rainer, and B. Regnell, *Case study research in software engineering: Guidelines and examples*: John Wiley & Sons, 2012.

[32] H. Sharp, Y. Dittrich, and C. R. B. de Souza, "The role of ethnographic studies in empirical software engineering," *IEEE Transactions on Software Engineering,* vol. 42, pp. 786-804, 2016.

[33] J. Rubin and D. Chisnell, *Handbook of usability testing: how to plan, design and conduct effective tests*: John Wiley & Sons, 2008.

[34] J. Nielsen, *Usability engineering*: Elsevier, 1994.
[35] J. S. Dumas and J. Redish, *A practical guide to usability testing*: Intellect Books, 1999.
[36] G. Salvendy, *Handbook of human factors and ergonomics*: John Wiley & Sons, 2012.
[37] J. R. Lewis, "Usability Testing," in *Handbook of human factors and ergonomics*, G. Salvendy, Ed., ed: John Wiley & Sons, Inc, 2012.
[38] B. Gillham, *The case study handbook*: Harvard Business School Boston, MA, 2007.
[39] J. A. Maxwell, "Designing a qualitative study," *The SAGE handbook of applied social research methods,* vol. 2, pp. 214-253, 2008.
[40] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," in *Guide to advanced empirical software engineering*, ed: Springer, 2008, pp. 285-311.
[41] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical software engineering,* vol. 14, pp. 131-164, 2009.
[42] C. B. Seaman, "Qualitative Methods in Empirical Studies of Software Engineering," *IEEE Transactions on Software Engineering,* vol. 25, pp. 557-572, 1999.
[43] T. C. Lethbridge, S. E. Sim, and J. Singer, "Studying software engineers: Data collection techniques for software field studies," *Empirical software engineering,* vol. 10, pp. 311-341, 2005.
[44] M. Van Den Haak, M. De Jong, and P. Jan Schellens, "Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue," *Behaviour & information technology,* vol. 22, pp. 339-351, 2003.
[45] J. Kontio, J. Bragge, and L. Lehtola, "The focus group method as an empirical tool in software engineering," in *Guide to advanced empirical software engineering*, ed: Springer, 2008, pp. 93-116.
[46] H. Priest, P. Roberts, and L. Woods, "An overview of three different approaches to the interpretation of qualitative data. Part 1: Theoretical issues," *Nurse Researcher (through 2013),* vol. 10, p. 43, 2002.
[47] L. Woods, H. Priest, and P. Roberts, "An overview of three different approaches to the interpretation of qualitative data. Part 2: practical illustrations," *Nurse Researcher (through 2013),* vol. 10, p. 43, 2002.
[48] C. Robson and K. McCartan, *Real world research*: Wiley, 2016.
[49] C. Andersson and P. Runeson, "A spiral process model for case studies on software quality monitoring—method and metrics," *Software Process: Improvement and Practice,* vol. 12, pp. 125-140, 2007.
[50] B. Kitchenham, L. Pickard, and S. L. Pfleeger, "Case studies for method and tool evaluation," *IEEE software,* vol. 12, p. 52, 1995.
[51] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam*, et al.*, "Preliminary Guidelines for Empirical Research in Software Engineering," *IEEE Transactions on Software Engineering,* vol. 28, pp. 721-734, 2002.
[52] C. Wohlin, M. Höst, and K. Henningsson, "Empirical research methods in software engineering," in *Empirical Methods and Studies in Software Engineering: Experiences from ESERNET*, ed: Springer, 2003, pp. 7-23.
[53] B. Kitchenham, H. Al-Khilidar, M. A. Babar, M. Berry, K. Cox, J. Keung*, et al.*, "Evaluating guidelines for reporting empirical software engineering studies," *Empirical Software Engineering,* vol. 13, pp. 97-121, 2008.
[54] D. E. Perry, S. E. Sim, and S. M. Easterbrook, "Case studies for software engineers," in *NASA SW Engineering Workshop Tutorial* 2005, pp. 736-738.
[55] R. K. Yin, *Case study research: Design and methods*: Sage publications, 2013.
[56] B. Kitchenham and e. al., "Guidelines for performing systematic literature reviews in software engineering," Keele University Keele, Staffs, ST5 5BG, UK Technical Report EBSE-2007-01, 9 July 2007 2007.
[57] M. Kuhrmann, D. M. Fernández, and M. Daneva, "On the pragmatic design of literature studies in software engineering: an experience-based guideline," *Empirical software engineering,* vol. 22, pp. 2852-2891, 2017.
[58] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from Applying the Systematic Literature Review Process within the Software Engineering domain," *Journal of Systems and Software,* vol. 80, pp. 571-583, 2007.
[59] R. K. Yin, *Case study research: Design and methods*, 2nd ed.: Sage publications, 1994.
[60] D. M. Fernández and J.-H. Passoth, "Empirical Software Engineering: From Discipline to Interdiscipline," *arXiv preprint arXiv:1805.08302,* 2018.
[61] F. Q. B. da Silva, M. Suassuna, A. C. C. França, A. M. Grubb, T. B. Gouveia, C. V. F. Monteiro*, et al.*, "Replication of empirical studies in software engineering research: a systematic mapping study," *Empirical Software Engineering,* vol. 19, pp. 501-557, 2014// 2014.
[62] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering*, 2008.
[63] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology,* vol. 64, pp. 1-18, 2015.
[64] Y. Zhou, H. Zhang, X. Huang, S. Yang, M. A. Babar, and H. Tang, "Quality assessment of systematic reviews in software engineering: a tertiary study," in *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, 2015, p. 14.
[65] M. Kuhrmann, C. Konopka, P. Nellemann, P. Diebold, and J. Münch, "Software process improvement: where is the evidence?: initial findings from a systematic mapping study," in *Proceedings of the 2015 International Conference on Software and System Process*, 2015, pp. 107-116.

[66] C. Wohlin, P. Runeson, P. A. d. M. S. Neto, E. Engström, I. do Carmo Machado, and E. S. De Almeida, "On the reliability of mapping studies in software engineering," *Journal of Systems and Software,* vol. 86, pp. 2594-2610, 2013.

[67] D. I. K. Sjoeberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N. K. Liborg*, et al.*, "A survey of controlled experiments in software engineering," *IEEE Transactions on Software Engineering,* vol. 31, pp. 733-753, 2005.