Methodology Matters: Is There a Method Choice Bias in Software Engineering?

Courtney Bornholdt
University of Victoria
BC, Canada
courtneywilliams@uvic.ca

Alexey Zagalsky
University of Victoria
BC, Canada
alexeyza@uvic.ca

Margaret-Anne Storey
University of Victoria
BC, Canada
mstorey@uvic.ca

ABSTRACT

As software engineering is a socio-technical research field, there are a myriad of research methods that researchers need to consider. Method choice determines different tradeoffs in terms of generalizabity, realism and control, among other attributes. In this paper, we reflect on the methods which tend to dominate software engineering research, and consider how humans tend to be involved in our research. We consider two years of ICSE proceedings and find that a majority of studies use computer simulation methods relying on trace measures rather than active human participation. This choice leads to a low level of control over extraneous factors. We question whether our method choice is out of convenience, or whether this choice reflects the kinds of research results our community values.

CCS CONCEPTS

• **General and reference** → *Empirical studies*;

KEYWORDS

Research Methods, Methodology, Human Involvement, Software Engineering

1 INTRODUCTION

Software engineering (SE) is often at the forefront of innovation and research, and involves both consideration of social and technical issues. Thus, we often employ a variety of methods in our studies [1, 2], but the choice of methods is a recurring topic of debate in our research community. Should certain methods be preferred over others? Are we striving for practical relevance, realism, or precision and control? The choice of methodology matters because it greatly impacts a study's advantages and its limitations.

One way to conceptualize our research is by categorizing papers as involving quantitative, qualitative, or mixed methods studies. While this may useful, data type is often conflated with research method. We suggest that this kind of classification is a naive choice for analysis because it fails to show the implications of method choice on aspects of research quality. In fact, there is a lack of awareness among researchers on how maximizing some desirable criteria may minimize others [6]. Instead, we select McGrath's eight core research strategies [3] as a way to classify software engineering research. McGrath highlights how methods can maximize certain dimensions while sacrificing others (universality, obtrusiveness, generalizability, realism, and control). We are also inspired by how he classifies data collection methods according to human participant involvement in the research process.

In this paper, we aim to reflect on and stimulate a discussion about research method choice and human involvement in SE and the impact it may have on generalizability, realism, and control. Since software engineering has a long tradition of using conference publications as the primary unit of dissemination, we choose the International Conference on Software Engineering (ICSE) as a our data source. The research questions that guide us are:

RQ1: What research and data collection methods have been used in ICSE publications?
RQ2: What is the "balance" between generalizability, control, and realism in research work published at ICSE?

2 METHODOLOGY

We conducted a systematic mapping study [4, 5]. First, we defined our research questions and collected relevant papers. We included all technical research track papers from ICSE's 2016 and 2017 proceedings in our study. We chose ICSE because it is a flagship conference in SE. We select two years as method preference may be influenced by a particular program committee. 101 and 68 technical track papers were analyzed from ICSE 2016 and 2017, respectively, for a total of 169 papers.

Second, we developed rules to use for our classification. We iterative refined McGrath's descriptions of data collection methods and research strategies as we applied them to our collection of SE papers. Through discussions, we adapted our interpretation of his descriptions. Our final description of McGrath's research strategies when applied to the SE domain are presented in Section 3.

Finally, the first author of this paper classified the papers according to these descriptions and recorded the classification in a spreadsheet.[1] To ensure valid and reliable outcomes, we used an independent researcher as part of an inter-rater agreement process: we provided this external researcher with our descriptions and a set of 34 (20%) randomly selected papers from the collected papers, which were independently classified to calculate a consensus estimate. As a result, we found a 72.1% consensus, which is within the 70% inter-rater threshold for consensus estimate quality [7].

Technical report DCS-358-IR, University of Victoria

[1] We publicly share the spreadsheet that includes the classification of our data on Zenodo—the link was removed for review purposes.

3 MCGRATH'S RESEARCH STRATEGIES

McGrath [3] proposed a set of eight research strategies in the form of a circumplex (see Fig. 1), positioned along two dimensions: the degree to which the setting used in the strategy is *universal* vs. *particular*; and the degree to which the strategy involves procedures that are *obtrusive* vs. *unobtrusive* with respect to the human systems under study. The circumplex also included three desired criteria (*generalizability*, *control*[2], and *realism*) and where each of the three is at its maximum. *Generalizability* refers to how generalizable the findings are to the population outside of the specific actors under study. *Realism* is how closely the context under which evidence is gathered matches real life. *Control* is defined as having control of the measurement of behaviors under study, as well as any extraneous factors not under study.

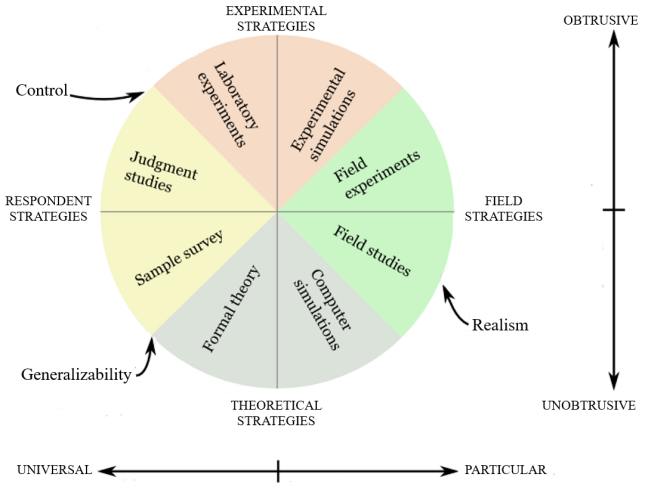

Figure 1: McGrath's research method circumplex.

McGrath described research strategies he used in behavioral and social sciences, not SE. However, the focus on human involvement is exactly why we believe that his research strategies would make an excellent and fitting *methodological lens* to gain insights on method choice in SE. After all, SE is a socio-technical field, one that deals with social aspects and is influenced by human behavior. Thus, we propose the following interpretations of McGrath's definitions of research strategies and data collection methods in order to accommodate for the socio-technical nature of SE.

3.1 Field Strategies

Field strategies in SE involve researchers entering the natural setting of the item under study in order to conduct their research. For example, this can be places where software development is occurring in action, such as a software company's offices. The distinction between a *Field Study* and a *Field Experiment* is the degree of *control* the researcher exercises in the situation under study.

In **field studies**, the researcher does not manipulate the setting and instead conducts their research using the "natural" environment setting. Observational studies are common types of field studies in SE. For example, a researcher visits a company that plans to adopt agile development and conducts a participant-observation study on how the shift to agile affects the employees and their roles.

Field experiments differ by introducing a controlled condition into the situation under study to understand the effects it creates—compromising some *unobtrusiveness* for higher *control* in the resulting study. Field experiments may be less common than field studies in SE, as industry participants may be unwilling to risk researchers introducing new interventions in their normal operating environment due to productivity or ethical concerns. An example of a SE field experiment can be introducing a novel automatic testing tool in a company and observing its effects on code quality.

3.2 Experimental Strategies

Experimental strategies in SE involve testing hypotheses in highly controlled situations. These strategies yield *high control in the measurements* and *control over extraneous factors* but at the cost of *reduced realism of context* and *narrowed generalizability*.

Laboratory experiments refer to situations created by the researchers, typically in their institutions, where individual participants or groups take part in an experiment. This strategy is used when researchers *focus on a certain behavior* and wish to measure it with *considerable control*. For example, a researcher investigating the effects of a new debugging tool on programming task efficiency may invite graduate students to a lab and ask them to accomplish a set of *predetermined* debugging tasks with and without the tool.

Experimental simulations in SE aim to *replicate some aspect of the participant's natural environment* during a controlled experiment, thus *gaining some realism*. For example, a researcher investigating project management meetings may conduct an experiment in a room with a similar setup to the one used at the company.

3.3 Respondent Strategies

Respondent strategies are used to systematically gather participant responses to questions posed by the researcher. The main difference between *Sample Surveys* and *Judgment Studies* is whether the study aims to gather *information about the human behavior under a stimulus* (i.e., respondent attributes), or *information about the stimulus itself*. These strategies make the participant's physical setting and conditions irrelevant.

Sample surveys in SE are used to investigate the effects that a phenomenon has on human behavior by surveying specific members of a chosen population, aiming at generalizing the findings to more of the population. For example, a researcher aiming to improve continuous integration tools may distribute an online survey, asking developers to describe how they use these tools and what challenges they face. Sample surveys are not limited to surveys in the traditional sense, and this method could also refer to interviews and focus groups. Sample surveys can be more convenient than field strategies because they often do not require physical access to an industrial environment and can be remotely conducted.

Judgment studies are commonly used in SE to evaluate the performance or utility of a new tool or technique. For example, in order to evaluate an API recommendation system, a researcher may invite developers to use the system and then survey them on

[2]In his paper, McGrath refers uses the terms "precision" and "control" synonymously. For the purpose of clarity, we use the term "control".

the relevance and accuracy of the resulting recommended APIs. Judgment studies tend to be *high on control of measurement* of both the stimulus materials and the responses; however, they are often *low on generalizability* of population, as they are done with "actors of convenience" or relatively small population samples.

3.4 Theoretical Strategies

Theoretical strategies differ from the previously described strategies because they are *the only methods that do not involve the inclusion of active human participants as part of the research* (but the studies may be based on past empirical work).

Computer simulations refer to controlled computer experiments that have a *complete and closed system* to model operations without any human involvement. These are common in SE and are typically used to evaluate the performance of an algorithm or technique based on existing and often publicly available data. Computer simulations can also be used for the development and verification of software and hardware systems. For example, a researcher aiming to evaluate a new bug detection technique may use version control history in an open-source project to see if their tool identified all the bugs that were fixed in subsequent versions of the project. Another example is running a series of experiments comparing the performance of various state-of-the-art static Android security analysis tools. Computer simulations may use methods for gathering and analyzing digitized data, which is common in data mining studies.

Formal theory research does not involve gathering new empirical data but rather focuses on the creation of models and theories based on previously gathered data or existing theories and models. For example, by building on a previously formed model, a theory formulation study may aim to identify and describe underlying factors, which can explain why certain practices support alignment and coordination in software projects.

3.5 Levels of Participant Involvement in Data Collection

McGrath discusses how data collection methods can be classified by type of human involvement. We use this to understand how SE researchers use human participants in their research.

Self reports refer to study instances where participants voluntarily report on their own behavior for research purposes. **Observations by a visible observer** and **observations by a hidden observer** are observations of human participants; either they are aware they are being observed or measured (visible observer) or they are unaware they are being observed (hidden observer). **Public archival records** and **private archival records** are records of human behavior that are recorded by a third party for non-research purposes, but are used as the subject of research after the fact. The difference between them is that private records would be unlikely to become a matter of public record, like a diary entry. The last method of data collection is **Trace Measures**, which refers to records indirectly created by humans as a result of their behavior. For example, software is written by developers to fulfill some need, but later the source code (or its bugs, commits, or error logs) becomes a trace measure we can study in future research. Self reports and observations are considered *active forms of human participation*, while archival records and trace measures are not.

Table 1: Method use in ICSE 2016 and 2017 papers

Quadrant	Method	'16	'17	Total
Field Strategies	Field Study	3	4	7 (3.6%)
	Field Experiment	4	2	6 (3.1%)
Experimental Strategies	Exp. Simulation	0	0	0 (0%)
	Lab Experiment	7	6	13 (6.6%)
Respondent Strategies	Judgment Study	4	7	11 (5.6%)
	Sample Survey	12	5	17 (8.7%)
Theoretical Strategies	Formal Theory	7	3	10 (5.1%)
	Comp. Simulation	78	54	132 (67.3%)

4 FINDINGS

Here, we report the findings of our mapping study. Some papers reported research that used multiple studies with different methods. In such cases, the paper was classified under all research methods used. In total, we recorded 196 research strategies and 215 data collection methods.

4.1 Method Choice in SE (RQ1)

We see a dramatic distinction between the use of computer simulations and all other methods (see Table 1): computer simulations were reported in 78 of ICSE 2016 and 54 of ICSE 2017 papers for a total of 67.3% of all methods used, where the next highest was 17 sample surveys for a total of 8.7%. The computer simulation designs were diverse ranging from data mining studies, computerized analysis of software artifacts, evaluation experiments of tools using software repositories and datasets of code artifacts, and computerized prediction and classification models. On the other extreme, no papers reported using experimental simulations.

Complementary to this, we classified papers by the level of human involvement (see Sec. 3.5). Figure 2 shows this classification of papers and reveals a big difference between the use of trace measures and other methods. When grouped by active human participation and no active human participation, we see that in total there are 62 studies (28.8%) that rely on active human participation methods, compared to 153 studies (71.2%) that rely on the use of traces and records of human behavior.

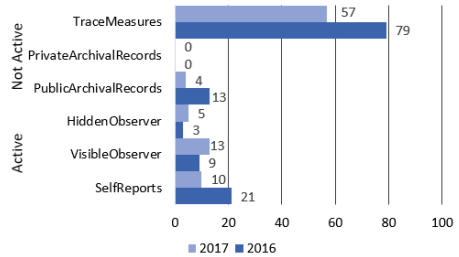

Figure 2: Human involvement in ICSE 2016 and 2017

4.2 How We Balance Generalizability, Realism, and Control in SE (RQ2)

To answer this question, we populated the quadrants of the circumflex based on the percentage of research strategies used in the papers we analyzed (see Fig. 3). Due to the high number of computer simulations, our results show a skew towards methods that achieve relatively high levels of realism and generalizability, but low levels of control.

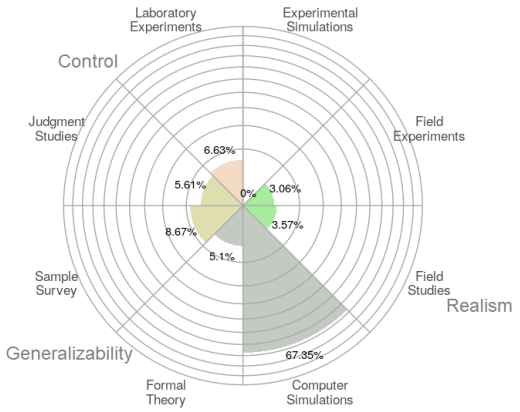

Figure 3: Research strategy choice ICSE 2016/2017

McGrath [3] also stipulates that *"to gain knowledge with confidence requires that more than one strategy—carefully selected so as to complement each other in strengths and weaknesses—be used in relation to any given problem"*. We found that 26 (15%) of the papers analyzed made use of multiple strategies. Some of these strategies were complementary; for example, we found 8 papers that included both a computer simulation and a judgment study. In these papers, the high control of the judgment studies makes up for the lack of control in the computer simulations, and likewise, the computer simulations make up for some of the lack of realism in the judgment studies. This is just one combination that allows researchers to triangulate their method use, allowing for higher amounts of all three desirable research criteria.

Plotting the use of research strategies allows us to show the balance of *universality* and *obtrusion* that may result from the choice of methods in our community. The high use of computer simulations leads to a low level of obtrusion. We also see a potential bias towards using particular and unobtrusive methods in research as opposed to universal or obtrusive methods. We discuss possible implications of this choice next.

5 DISCUSSION

From our findings, we see an **imbalance in method choice** (in particular much more use of computer simulations over other methods). This mono-method choice may lead to a poor balance of achieved generalizability, realism, and control if we consider all studies across our research community. In particular, control seems to suffer from our choice of methods, but realism and generalizability may also suffer if we seldom use methods that can maximize those criteria. As a community, we should discuss if more triangulation of methods (at a community level - if not study level) is a desirable goal to strive for, but also discuss why these other methods, which can bring valuable insights in socio-technical fields, are not used more.

We also found that the studies that are reported at ICSE tend to rely on **trace measures and archival records over other classes of data**. Trace measures, as well as archival records, have the advantage of being unobtrusive and easy to collect, and such data is not biased by knowledge that the data will be used for research purposes. However, there is often only a loose link between the record and the concept it is being used for[3]. In particular, we cannot always be sure that we are measuring and studying what we intend to measure because of factors beyond our knowledge and control. When we consider that most of the trace measures were used as data for computer simulations, this means we see even lower control over extraneous variables.

Direct human participation allows us to control for extraneous variables because we can ask participants themselves about what is occurring in their world and why. McGrath discusses that similar to the research methods, the different classes of data collection have strengths and weaknesses, but can be used in combination with each other to compensate for these weaknesses. While it may be easier to use trace methods in research than to actively engage humans, we propose that we discuss as a community whether it is in our best interests to include more active human participation to enrich our body of work in the future.

McGrath's proposed his circumflex before we had today's rapid advancements in social technologies and increasing availability of big data. The advancements offer new opportunities for extending and amplifying existing research methods (e.g., lower cost, increased reach), while these advancements may also mitigate existing limitations (e.g., response rate, representativeness of population). Thus, it is not surprising that virtual research methods are gaining popularity. However, it is important to consider how these advancements affect generalizability, control, and realism. When is it appropriate to substitute in-person interaction with technological interactions? Furthermore, ethical concerns should be revisited and discussed, as these advancements bring new pitfalls as well, such as loss of contextual information surrounding recorded data.

We hope that the results we report in this paper will spark some discussion on the choice of research and data collection methods, and its impact on achieving desirable research criteria in software engineering.

REFERENCES

[1] John W Creswell. 2013. *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications.

[2] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg. 2002. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering* 28, 8 (Aug 2002), 721–734.

[3] Joseph E. Mcgrath. 1995. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd ed)*. Citeseer.

[4] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic Mapping Studies in Software Engineering.. In *EASE*, Vol. 8. 68–77.

[5] Mary Shaw. 2003. Writing good software engineering research papers. In *Software Engineering, 2003. Proceedings. 25th International Conference on.* IEEE, 726–736.

[6] Janet Siegmund, Norbert Siegmund, and Sven Apel. 2015. Views on internal and external validity in empirical software engineering. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, Vol. 1. IEEE, 9–19.

[7] Steven E Stemler. 2004. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation* 9, 4 (2004), 1–19.