ICS 500: Research Methods and Experiment Design in Computing

Lecture

Empirical Strategies

The slides are prepared by: Dr. Mahmood Niazi, Dr. Malak Baslyman, Dr. Hamoud Aljamaan & Dr. Mohammad Alshayeb

Lecture Objectives

- ✓ Introduce empirical research strategies
- ✓ Highlight important aspects in relation to the empirical strategies
- ✓ Illustrate how the strategies can be used

Empirical Research

- The American Heritage Dictionary of the English Language defines **empirical** as:
  - Relying on or derived from observation or experiment, verifiable or provable by means of observation or experiment and guided by **practical experience and not theory.**

- The Macquarie Dictionary defines **empirical** as:
  - Taken from or guided by experience or experiment.
  - Depending upon experience or observation rather than using theory.

- These definitions tell us that empirical research is the type of research which is based on **observed** and **measured phenomena** that derives knowledge from experience rather than from theory.

Research Paradigms

- There are two types of research paradigms that have different approaches to empirical studies:
  - Exploratory
  - Explanatory

Exploratory Research

- Conducted for a problem that has not been clearly defined and it is used to gather preliminary information that will help define problems and suggest hypotheses/ research questions.

- This implies that a **flexible** research design is needed to adapt to changes in the observed phenomenon.

- Flexible design research is also referred to as qualitative research, as it primarily is informed by qualitative data.

- It is concerned with discovering causes noticed by the subjects in the study and understanding their view of the problem at hand.

Exploratory Research Examples

- Our sales are declining, and we do not know why
- Would people be interested in our new product idea?
- Would training improve programmers' capabilities?

Explanatory Research

- Mainly concerned with quantifying a relationship or to compare two or more groups with the aim to identify a cause-effect relationship (researchers want to explain what is going on).
- This type of study is a fixed design study, implying that factors are fixed before the study is launched.
- Fixed design research is also referred to as quantitative research, as it primarily is informed by quantitative data.
- An advantage is that quantitative data promotes comparisons and statistical analyses.

Explanatory Research Examples

- Which of two testing methods is more effective in terms of detecting bugs?
- Which of the two-programming language is easier to learn?
- Which of two ML algorithms is more effective in terms of detecting type 1 diabeties?

Research Paradigms *(cont'd)*

- It is possible for qualitative and quantitative research to investigate the same topics but each of them will address a different type of question.

- For example
  - How much does a new inspection method decrease the number of faults found in test?
    - (Quantitative investigation)
  - What are the sources of variations between different inspection groups?
    - (Qualitative investigation)

Research Paradigms *(cont'd)*

- **Fixed** design strategies, such as controlled experiments, are appropriate when testing the effects of a treatment.
  - In experiments, a ***treatment*** is something that researchers administer to experimental units. For example, a doctor treats a patient with a skin condition with different creams to see which is most effective.
  - A teacher practices different teaching methods on different groups in his/her class to see which yields the best results

- A **flexible** design strategies are appropriate to find out why the results from a quantitative investigation are as they are.

- The two approaches should be regarded as complementary rather than competitive.

Empirical Strategies/Methods

- Depending on the purpose of the evaluation, whether it is techniques, methods or tools, and depending on the conditions for the empirical investigation, there are three major strategies that may be carried out:
  - Survey
  - Case Study
  - Experiment

Survey

- A **survey** is a system for collecting information from or about people to describe, compare or explain their knowledge, attitudes and behavior
- A survey is often an investigation performed in retrospect (a past course of events), when, for example, a tool or technique, has been in use for a while.
- The primary means of gathering qualitative or quantitative data are interviews or questionnaires. These are done through taking a sample which is representative from the population to be studied.
- The results from the survey are then **analyzed** to derive descriptive and explanatory conclusions.

Survey *(cont'd)*

- Example: studying how a new development process has improved the developers' attitudes towards quality assurance.
  1. A sample of developers is selected from all the developers at the company.
  2. A questionnaire is constructed to obtain information needed for the research.
  3. The questionnaires are answered by the sample of developers.
  4. The information collected are then arranged into a form that can be handled in a quantitative or qualitative manner.

Survey Characteristics

- Surveys are not conducted to create an understanding of the particular sample. Instead, the purpose is to **understand the population**, from which the sample was drawn.
  - For example, by interviewing 25 developers on what they think about a new process, the opinion of the larger population of 200 developers in the company can be assessed.
  - Surveys aim at the development of generalized conclusions.

Survey Purposes

- **Descriptive surveys** can be conducted to enable assertions (claims or declaration) about some population.
  - For example, 50% of the developers use C++
- **Explanatory surveys** aim at making explanatory claims about the population.
  - For example, we want to explain why some developers prefer one technique while others prefer another.
- **Explorative surveys** provide new possibilities that could be further analyzed and followed up in the more focused thorough survey.
  - For example, we want to explore reasons of bugs in software code.

Survey Data collection

- The mechanism used for data collection is usually termed the instrument

- In software engineering, surveys are commonly conducted by using the following two forms of instrument:
  - **Questionnaires**. The act of completing a questionnaire, either on paper or on-line is often thought of as almost synonymous with the idea of a survey.
  - **Interviews**. They can be more tightly targeted at a particular group, and if using semi-structured interviews, it may also be possible to probe more deeply into the issues identified (usually related to "why").
    - Interviews are time-consuming to perform, as well as requiring the researcher to possess some 'people skills', and it may also be more difficult to access the relevant group of respondents.

Survey Instruments

- Survey instruments usually make use of two forms of question:
  - Open questions do not have any pre-determined set of possible answers and are generally useful for collecting data for descriptive surveys.
    - They offer flexibility in proving issues, analysis of the responses
    - Need a qualitative approach.
  - Closed questions may ask the respondent to select one or more values from a pre-set list (a rating question), a form that includes the use of Likert scales, or to order a set of pre-determined options (a ranking question).
    - Can be used in a quantitative analysis.

Sampling

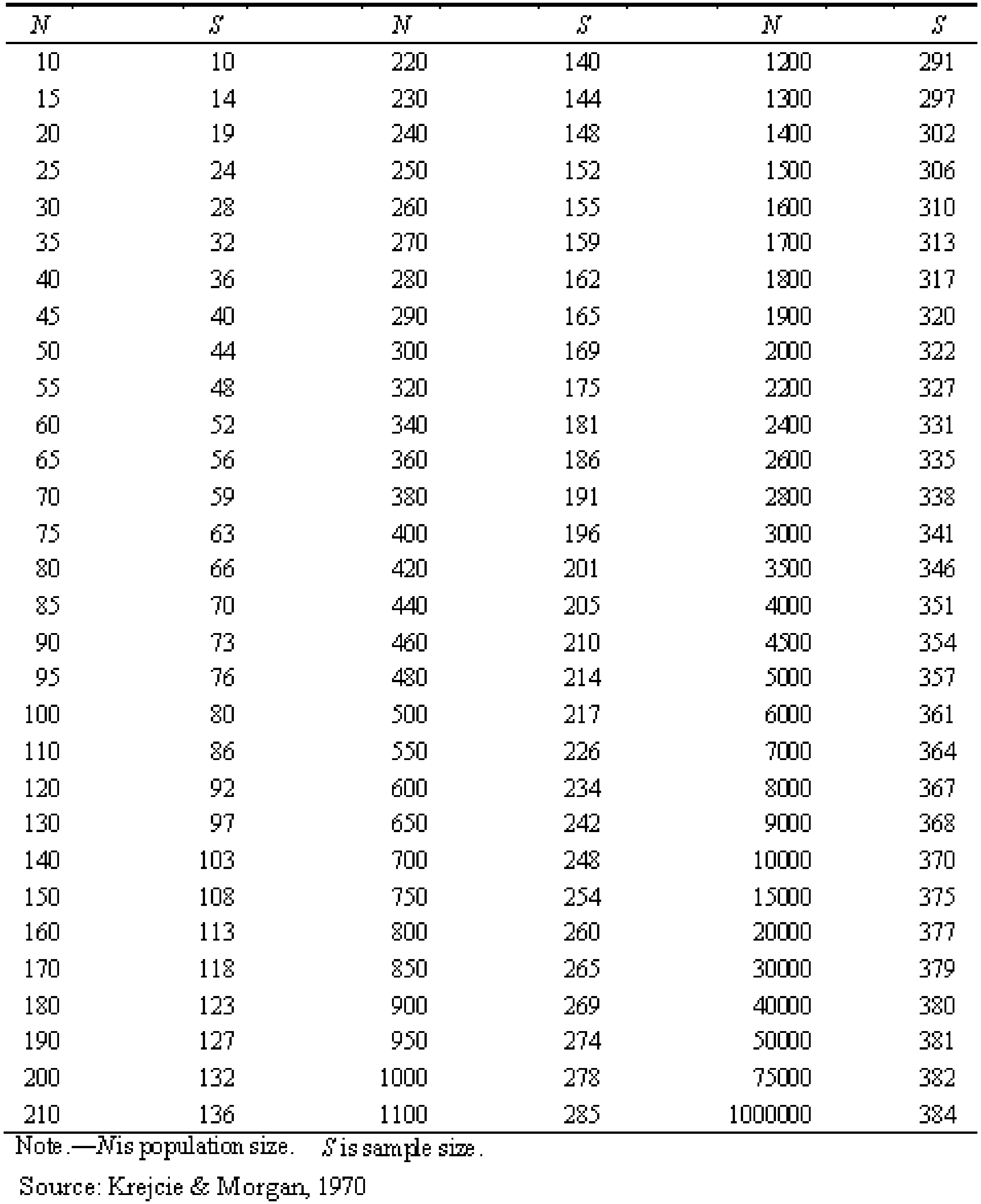
- Sampling is important from two aspects (sampling frame).
  - In planning a survey, it is necessary to decide what the population we are interested in
  - From which we wish to draw our respondents

- A sampling frame is a list of all those within a population who can be sampled
- For example, we might be interested in conducting a survey of everyone who has had experience with using pair programming on a software development project (sampling frame)
  - Defining exactly what is required for potential respondents to qualify for membership of this (in terms of measures that can be used to define some basic level of experience) is quite difficult - what exactly we mean by such words as "using".

Sample Size

- Define the size of sample that we need to obtain from in order to be able to make sound inferences about the larger group.
  - Inference: a conclusion reached on the basis of evidence, e.g., if you see someone eating a new food and he or she makes a face, then you infer he does not like it..
- If we do know, or can reasonably estimate, the size of the sampling frame, then it is possible to determine the size of sample that we need (the number of respondents) in order to achieve the required confidence interval in our results.
- As a ‘rule of thumb’, a minimum of at least 30 responses is really desirable if aiming to perform any form of statistical analysis.

Sample Size



N	S	N	S	N	S
10	10	220	140	1200	291
15	14	230	144	1300	297
20	19	240	148	1400	302
25	24	250	152	1500	306
30	28	260	155	1600	310
35	32	270	159	1700	313
40	36	280	162	1800	317
45	40	290	165	1900	320
50	44	300	169	2000	322
55	48	320	175	2200	327
60	52	340	181	2400	331
65	56	360	186	2600	335
70	59	380	191	2800	338
75	63	400	196	3000	341
80	66	420	201	3500	346
85	70	440	205	4000	351
90	73	460	210	4500	354
95	76	480	214	5000	357
100	80	500	217	6000	361
110	86	550	226	7000	364
120	92	600	234	8000	367
130	97	650	242	9000	368
140	103	700	248	10000	370
150	108	750	254	15000	375
160	113	800	260	20000	377
170	118	850	265	30000	379
180	123	900	269	40000	380
190	127	950	274	50000	381
200	132	1000	278	75000	382
210	136	1100	285	1000000	384

Note.—*N* is population size. *S* is sample size.

Source: Krejcie & Morgan, 1970

Krejcie, R.V., & Morgan, D.W., (1970). Determining Sample Size for Research Activities. Educational and Psychological Measurement.

Categories of sampling technique

- **Probabilistic Sampling**. Seeks to obtain a sample that is intended to be a representative cross-section of the population.
  - Depending on the nature of the sampling frame, a random sampling of the population may have a similar profile to the overall population.
  - For example, if years of experience is a key factor, then this might be used as the basis for the stratification (categorization).

- **Non-Probabilistic Sampling**. Forms a poorer basis for inference about the whole population, it may well be the only option available.
  - For example, we construct a website to collect the data and then post invitations to people to respond using this (termed self-selection sampling).
  - Other strategies include:
    - Purposive sampling (sending requests to people who have specific characteristics and are likely to respond)
    - Snowball sampling (asking respondents to identify others who might be willing to participate);
    - Convenience sampling where people are recruited on the basis of being readily available or likely to be willing to take part.

Collecting Survey Response

- Surveys tend to have low response rates, with a figure of 10% usually being considered to be quite good.
- Ways to improve this include providing a good explanation about why participation is important when making the original request and following up non-responders with a reminder after a suitable interval of time.
- With on-line surveys there may also be instances of partial completion, where people have started to enter their responses but have not completed the task.
  - Criteria for deciding when such responses should be included need to be specified when writing the research protocol for the survey.

Questionnaire Design

- Design of the questionnaire, or the set of questions to be used in a semi-structured interview is an important design element.
- Questions need to be clear, and to check consistency of responses, it may be appropriate to use more than one question to address a topic.
- The set of questions should be assessed for the following qualities.
  - Ensuring that the questions address a well-balanced sample of issues (content validity).
  - Measuring the relevant attributes through the data collected by the questions (construct validity).
  - Being confident that if the questions were given repeatedly to the same people that they would obtain the same answers (reliability).

Reporting Surveys

- Reporting many of the same issues as for experiments
  - Research question, design, conduct, results, analysis, outcomes, and threats to validity.
  - Where appropriate, there may also be hypotheses to report as well.
  - Information about issues as the population of interest, the sampling frame used, and the way that the sampling was performed.

Case Study

- A **case study** is conducted to investigate a single entity or phenomenon (fact) in its real-life context, within a specific time space.

- The researcher collects detailed information on, for example, one single project during some period of time.

- A case study can be applied as a comparative research strategy.
  - For example, you want to evaluate or compare that your proposed solution is better than the existing one in the real-world environment.

- Case studies are very suitable for industrial evaluation of software engineering methods and tools

Case Study Confounding Factors

- When performing case studies, it is necessary to minimize the effects of confounding factors.

- A confounding factor is a factor that makes it impossible to distinguish the effects of two factors from each other.
  - For example, it may be difficult to tell if a better result depends on the tool or the experience of the user of the tool.

Experiments

- Scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact.

- Experiment (or controlled experiment) is an empirical enquiry that manipulates one factor or variable of the studied setting.
  - Based on randomization, different treatments are applied to or by different subjects, while keeping other variables constant, and measuring the effects on outcome variables.

- Experiments are conducted when we want control over the situation and want to manipulate behavior directly, precisely and systematically.
  - For example, it is possible to control who is using one method and who is using another method

Experiments

- Experiments are mostly done in a laboratory environment, which provides a high level of control.
- When experimenting, subjects are assigned to different treatments at random.
- The objective is to manipulate one or more variables and control all other variables at fixed levels.
- The effect of the manipulation is measured and based on this a statistical analysis can be performed.
- In cases where it is impossible to randomly assign treatments to subjects, we may use **quasi-experiments**.

Experiments

- **Quasi-experiment** is an empirical enquiry that is similar to an experiment, where the assignment of treatments to subjects cannot be based on randomization, but emerge from the characteristics of the subjects or objects themselves, who is using one method and who is using another method

- For example
  - For ethical reasons, subjects must be allowed to chose their treatment.
  - If an experiment is performed in a company, there may be constraints on which employees can work on which tasks.

Quasi experiments

- Software engineering experiments are often quasi experiments, i.e., experiment in which it, for example, has not been possible to assign participants in the experiments to groups by random.

- In cases where it is impossible to randomly assign treatments to subjects, we may use quasi-experiments

Quasi-experiment example

- A researcher who wants to evaluate a new method of teaching fractions to third graders.
- One way would be to conduct a study with a treatment group consisting of one class of third-grade students and a control group consisting of another class of third-grade students.
- This design would be a nonequivalent groups design because the students are not randomly assigned to classes by the researcher, which means there could be important differences between them.
  - For example, the parents of higher achieving or more motivated students might have been more likely to request that their children be assigned to Mr. X class.
  - Or the principal might have assigned the "troublemakers" to Mr. Y's class because he is a stronger disciplinarian.
  - Of course, the teachers' styles, and even the classroom environments, might be very different and might cause different levels of achievement or motivation among the students.
- If at the end of the study there was a difference in the two classes' knowledge of fractions, it might have been caused by the difference between the teaching methods—but it might have been caused by any of these confounding variables.

Human-oriented vs. Technology-oriented Experiments

- In human-oriented experiments, humans apply different treatments to objects
  - For example, two inspection methods (treatments) are applied (by human) to one pieces of code (object).
- In technology-oriented experiments, typically different tools are applied to different objects,
  - For example, two test case generation tools (treatments) are applied to the same programs (object).
- The human-oriented experiment has less control than the technology-oriented one, since humans behave differently at different occasions, while tools (mostly) are deterministic.
  - Due to learning effects, a human subject cannot apply two methods to the same piece of code (as after first method human will learn and his application of second method will be different), which two tools can do without bias.

Replication in Experiments

- The replication of an experiment involves repeating the investigation under similar conditions, while for example, varying the subject population.
- The purpose of a replication is to show that the result from the original experiment is valid for a larger population.

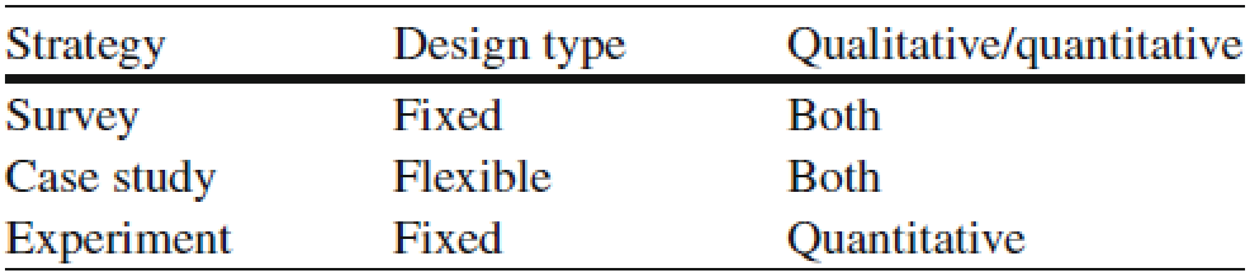
Case Studies Vs Experiments

- A case study is an observational study while the experiment is a controlled study.

- An advantage of case studies is that they are easier to plan and are more realistic.

- The disadvantages of case studies are that the results are difficult to generalize and harder to interpret.
  - It is possible to show the effects in a typical situation, but it requires more analysis to generalize to other situations

Empirical strategies comparison



Strategy	Design type	Qualitative/quantitative
Survey	Fixed	Both
Case study	Flexible	Both
Experiment	Fixed	Quantitative

Empirical strategies comparison

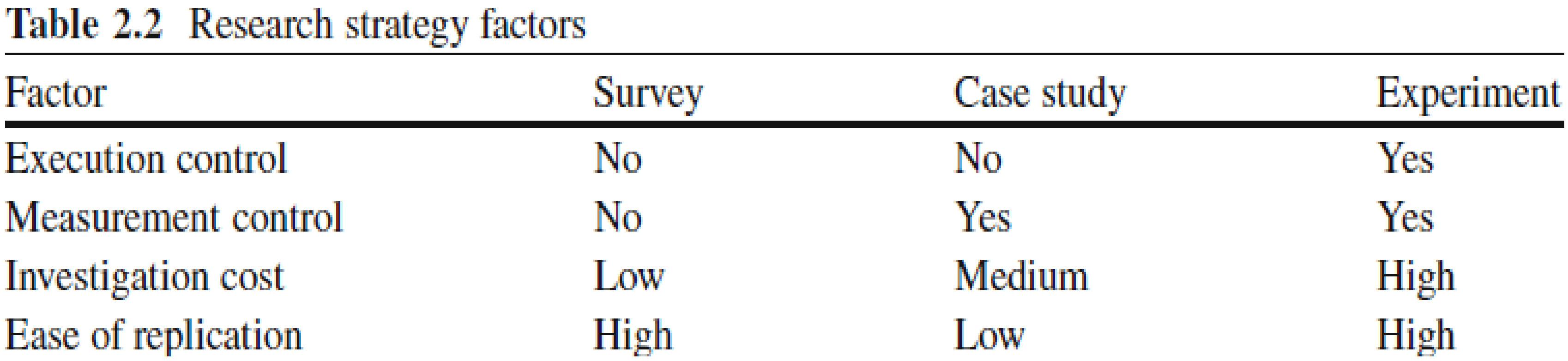


Table 2.2 Research strategy factors

Factor	Survey	Case study	Experiment
Execution control	No	No	Yes
Measurement control	No	Yes	Yes
Investigation cost	Low	Medium	High
Ease of replication	High	Low	High

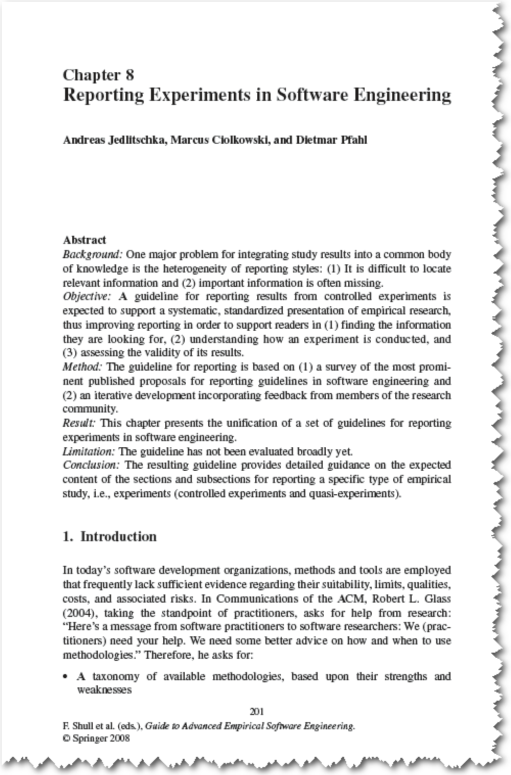
- Execution control describes how much control the researcher has over the study.

- Measurement control is the degree to which the researcher can decide upon which measures to be collected, and to include or exclude during execution of the study
  - An example is how to collect data about requirement volatility

Aggregating evidence from empirical studies

- The research should build upon each other so new research should always take existing knowledge into consideration as its starting point.
- Several empirical studies may together give answers to questions, which are not sufficiently answered by individual studies in isolation.
- Systematic literature reviews are means to collect and synthesize empirical evidence from different sources.

Reporting Experiments



Chapter 8
Reporting Experiments in Software Engineering

Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl

Abstract

Background: One major problem for integrating study results into a common body of knowledge is the heterogeneity of reporting styles: (1) It is difficult to locate relevant information and (2) important information is often missing.
Objective: A guideline for reporting results from controlled experiments is expected to support a systematic, standardized presentation of empirical research, thus improving reporting in order to support readers in (1) finding the information they are looking for, (2) understanding how an experiment is conducted, and (3) assessing the validity of its results.
Method: The guideline for reporting is based on (1) a survey of the most prominent published proposals for reporting guidelines in software engineering and (2) an iterative development incorporating feedback from members of the research community.
Result: This chapter presents the unification of a set of guidelines for reporting experiments in software engineering.
Limitation: The guideline has not been evaluated broadly yet.
Conclusion: The resulting guideline provides detailed guidance on the expected content of the sections and subsections for reporting a specific type of empirical study, i.e., experiments (controlled experiments and quasi-experiments).

1. Introduction

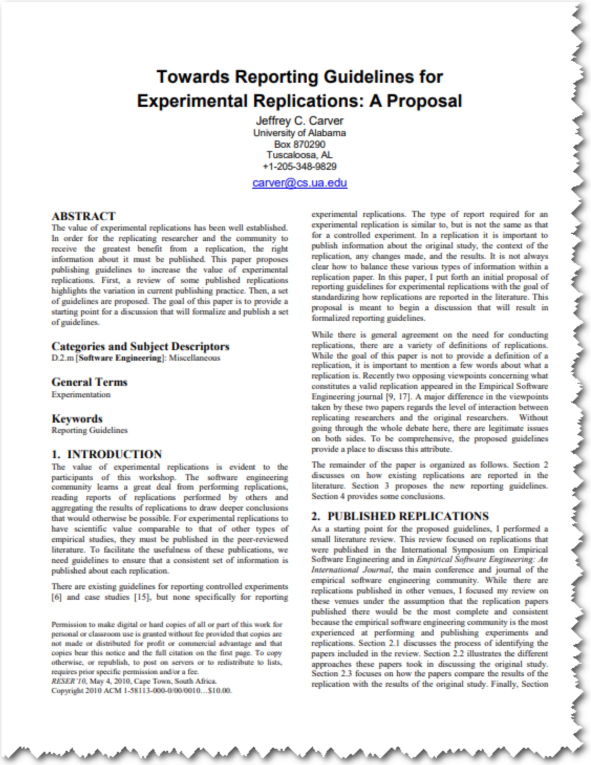
In today's software development organizations, methods and tools are employed that frequently lack sufficient evidence regarding their suitability, limits, qualities, costs, and associated risks. In Communications of the ACM, Robert L. Glass (2004), taking the standpoint of practitioners, asks for help from research: "Here's a message from software practitioners to software researchers: We (practitioners) need your help. We need some better advice on how and when to use methodologies." Therefore, he asks for:

- A taxonomy of available methodologies, based upon their strengths and weaknesses

201

F. Shull et al. (eds.), *Guide to Advanced Empirical Software Engineering.*
© Springer 2008

F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering, Springer 2008



Towards Reporting Guidelines for Experimental Replications: A Proposal

Jeffrey C. Carver
University of Alabama
Box 870290
Tuscaloosa, AL
+1-205-348-9829
carver@cs.ua.edu

ABSTRACT

The value of experimental replications has been well established. In order for the replicating researcher and the community to receive the greatest benefit from a replication, the right information about it must be published. This paper proposes publishing guidelines to increase the value of experimental replications. First, a review of some published replications highlights the variation in current publishing practice. Then, a set of guidelines are proposed. The goal of this paper is to provide a starting point for a discussion that will formalize and publish a set of guidelines.

Categories and Subject Descriptors

D.2.m [**Software Engineering**]: Miscellaneous

General Terms

Experimentation

Keywords

Reporting Guidelines

1. INTRODUCTION

The value of experimental replications is evident to the participants of this workshop. The software engineering community learns a great deal from performing replications, reading reports of replications performed by others and aggregating the results of replications to draw deeper conclusions that would otherwise be possible. For experimental replications to have scientific value comparable to that of other types of empirical studies, they must be published in the peer-reviewed literature. To facilitate the usefulness of these publications, we need guidelines to ensure that a consistent set of information is published about each replication.

There are existing guidelines for reporting controlled experiments [6] and case studies [15], but none specifically for reporting experimental replications. The type of report required for an experimental replication is similar to, but is not the same as that for a controlled experiment. In a replication it is important to publish information about the original study, the context of the replication, any changes made, and the results. It is not always clear how to balance these various types of information within a replication paper. In this paper, I put forth an initial proposal of reporting guidelines for experimental replications with the goal of standardizing how replications are reported in the literature. This proposal is meant to begin a discussion that will result in formalized reporting guidelines.

While there is general agreement on the need for conducting replications, there are a variety of definitions of replications. While the goal of this paper is not to provide a definition of a replication, it is important to mention a few words about what a replication is. Recently two opposing viewpoints concerning what constitutes a valid replication appeared in the Empirical Software Engineering journal [9, 17]. A major difference in the viewpoints taken by these two papers regards the level of interaction between replicating researchers and the original researchers. Without going through the whole debate here, there are legitimate issues on both sides. To be comprehensive, the proposed guidelines provide a place to discuss this attribute.

The remainder of the paper is organized as follows. Section 2 discusses on how existing replications are reported in the literature. Section 3 proposes the new reporting guidelines. Section 4 provides some conclusions.

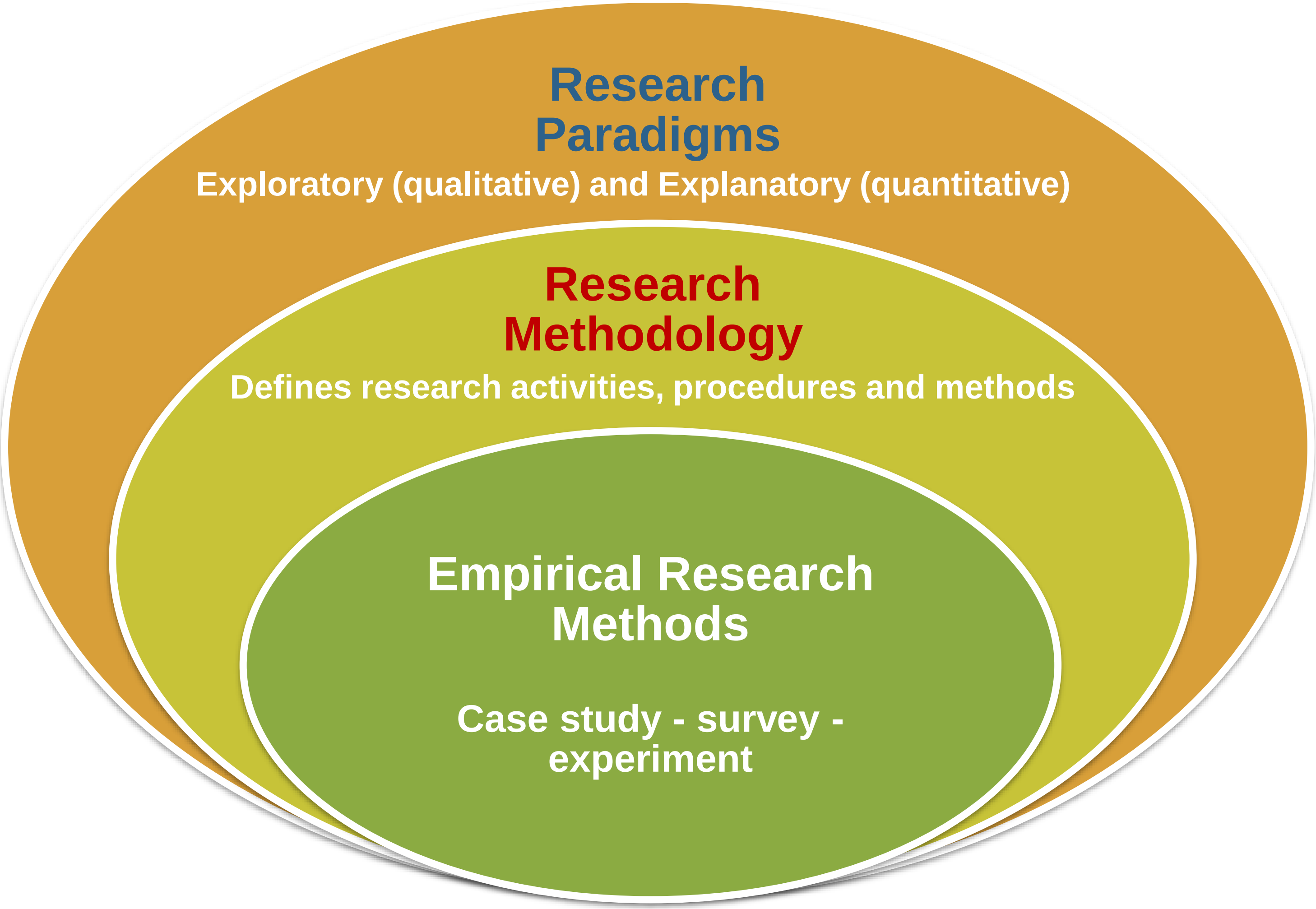
2. PUBLISHED REPLICATIONS

As a starting point for the proposed guidelines, I performed a small literature review. This review focused on replications that were published in the International Symposium on Empirical Software Engineering and in *Empirical Software Engineering: An International Journal*, the main conference and journal of the empirical software engineering community. While there are replications published in other venues, I focused my review on these venues under the assumption that the replication papers published there would be the most complete and consistent because the empirical software engineering community is the most experienced at performing and publishing experiments and replications. Section 2.1 discusses the process of identifying the papers included in the review. Section 2.2 illustrates the different approaches these papers took in discussing the original study. Section 2.3 focuses on how the papers compare the results of the replication with the results of the original study. Finally, Section

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
RESER'10, May 4, 2010, Cape Town, South Africa.
Copyright 2010 ACM 1-58113-000-0/00/0010…$10.00.

Carver, J., "Towards Reporting Guidelines for Experimental Replications: A Proposal." Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER) [Held during ICSE 2010]. May 4, 2010. Cape Town, South Africa.

Summary



Research Paradigms

Exploratory (qualitative) and Explanatory (quantitative)

Research Methodology

Defines research activities, procedures and methods

Empirical Research Methods

Case study - survey - experiment

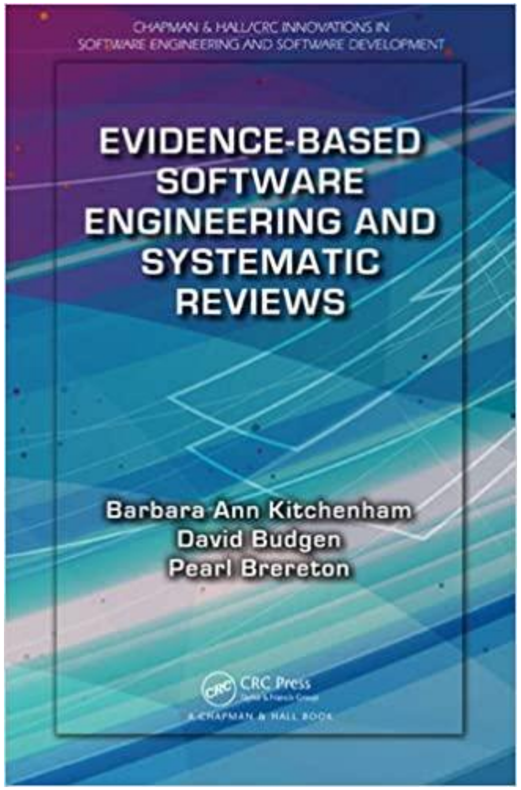
Resource (1/2)

Claes Wohlin · Per Runeson
Martin Höst · Magnus C. Ohlsson
Björn Regnell · Anders Wesslén

Experimentation in Software Engineering

Springer

Chapter 2



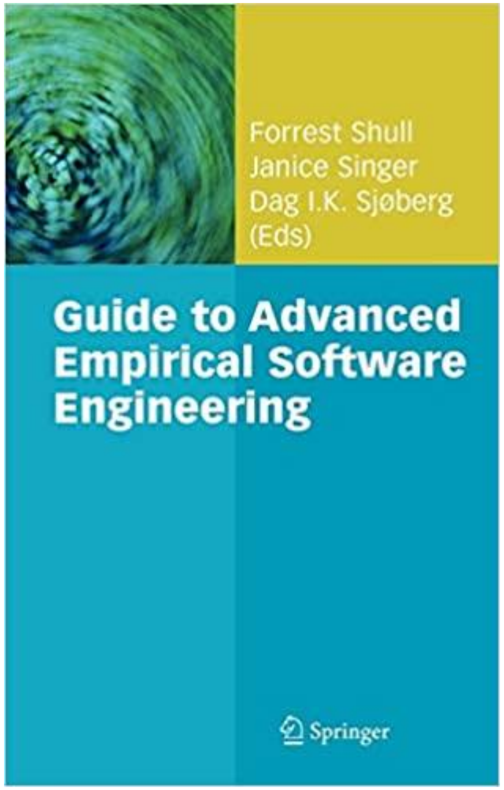
CHAPMAN & HALL/CRC INNOVATIONS IN SOFTWARE ENGINEERING AND SOFTWARE DEVELOPMENT

EVIDENCE-BASED SOFTWARE ENGINEERING AND SYSTEMATIC REVIEWS

Barbara Ann Kitchenham
David Budgen
Pearl Brereton

CRC Press
A CHAPMAN & HALL BOOK

Chapter 17



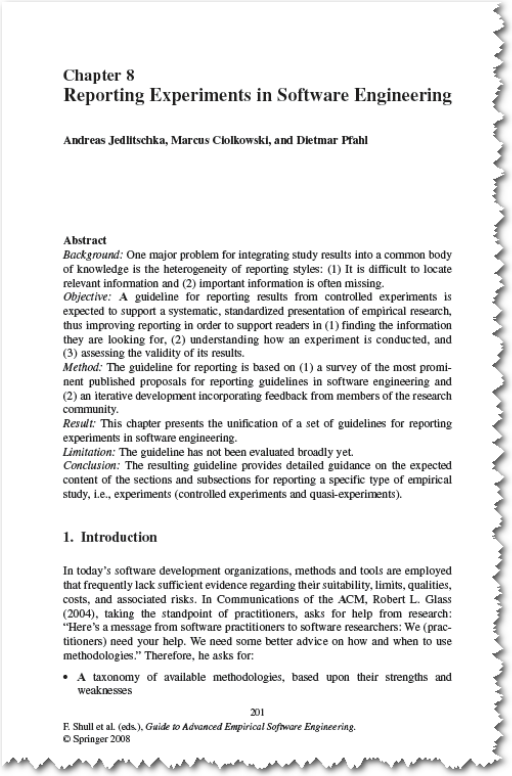
Forrest Shull
Janice Singer
Dag I.K. Sjøberg
(Eds)

Guide to Advanced Empirical Software Engineering

Springer

Chapter 3

Resource (2/2)



Chapter 8

Reporting Experiments in Software Engineering

Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl

Abstract

Background: One major problem for integrating study results into a common body of knowledge is the heterogeneity of reporting styles: (1) It is difficult to locate relevant information and (2) important information is often missing.
Objective: A guideline for reporting results from controlled experiments is expected to support a systematic, standardized presentation of empirical research, thus improving reporting in order to support readers in (1) finding the information they are looking for, (2) understanding how an experiment is conducted, and (3) assessing the validity of its results.
Method: The guideline for reporting is based on (1) a survey of the most prominent published proposals for reporting guidelines in software engineering and (2) an iterative development incorporating feedback from members of the research community.
Result: This chapter presents the unification of a set of guidelines for reporting experiments in software engineering.
Limitation: The guideline has not been evaluated broadly yet.
Conclusion: The resulting guideline provides detailed guidance on the expected content of the sections and subsections for reporting a specific type of empirical study, i.e., experiments (controlled experiments and quasi-experiments).

1. Introduction

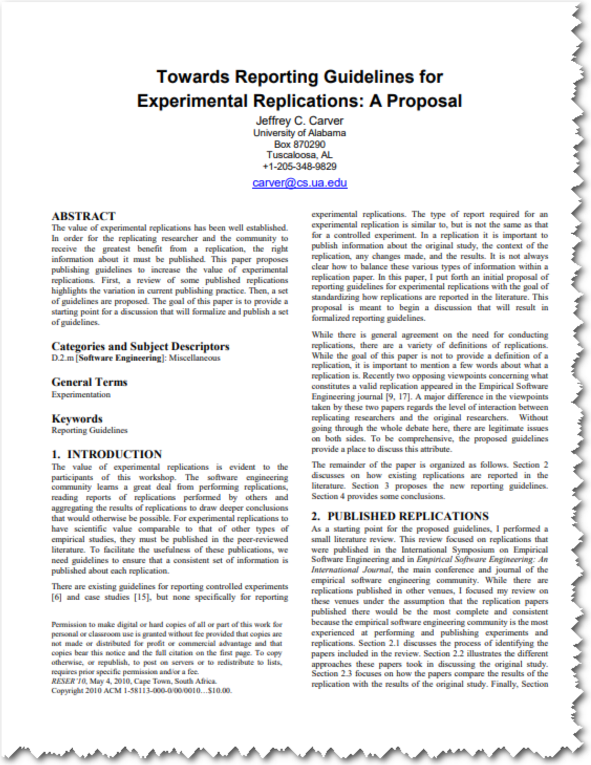
In today's software development organizations, methods and tools are employed that frequently lack sufficient evidence regarding their suitability, limits, qualities, costs, and associated risks. In Communications of the ACM, Robert L. Glass (2004), taking the standpoint of practitioners, asks for help from research: "Here's a message from software practitioners to software researchers: We (practitioners) need your help. We need some better advice on how and when to use methodologies." Therefore, he asks for:

- A taxonomy of available methodologies, based upon their strengths and weaknesses

201

F. Shull et al. (eds.), *Guide to Advanced Empirical Software Engineering.*
© Springer 2008

F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering, Springer 2008



Towards Reporting Guidelines for Experimental Replications: A Proposal

Jeffrey C. Carver
University of Alabama
Box 870290
Tuscaloosa, AL
+1-205-348-9829
carver@cs.ua.edu

ABSTRACT

The value of experimental replications has been well established. In order for the replicating researcher and the community to receive the greatest benefit from a replication, the right information about it must be published. This paper proposes publishing guidelines to increase the value of experimental replications. First, a review of some published replications highlights the variation in current publishing practice. Then, a set of guidelines are proposed. The goal of this paper is to provide a starting point for a discussion that will formalize and publish a set of guidelines.

Categories and Subject Descriptors

D.2.m [**Software Engineering**]: Miscellaneous

General Terms

Experimentation

Keywords

Reporting Guidelines

1. INTRODUCTION

The value of experimental replications is evident to the participants of this workshop. The software engineering community learns a great deal from performing replications, reading reports of replications performed by others and aggregating the results of replications to draw deeper conclusions that would otherwise be possible. For experimental replications to have scientific value comparable to that of other types of empirical studies, they must be published in the peer-reviewed literature. To facilitate the usefulness of these publications, we need guidelines to ensure that a consistent set of information is published about each replication.

There are existing guidelines for reporting controlled experiments [6] and case studies [15], but none specifically for reporting experimental replications. The type of report required for an experimental replication is similar to, but is not the same as that for a controlled experiment. In a replication it is important to publish information about the original study, the context of the replication, any changes made, and the results. It is not always clear how to balance these various types of information within a replication paper. In this paper, I put forth an initial proposal of reporting guidelines for experimental replications with the goal of standardizing how replications are reported in the literature. This proposal is meant to begin a discussion that will result in formalized reporting guidelines.

While there is general agreement on the need for conducting replications, there are a variety of definitions of replications. While the goal of this paper is not to provide a definition of a replication, it is important to mention a few words about what a replication is. Recently two opposing viewpoints concerning what constitutes a valid replication appeared in the Empirical Software Engineering journal [9, 17]. A major difference in the viewpoints taken by these two papers regards the level of interaction between replicating researchers and the original researchers. Without going through the whole debate here, there are legitimate issues on both sides. To be comprehensive, the proposed guidelines provide a place to discuss this attribute.

The remainder of the paper is organized as follows. Section 2 discusses on how existing replications are reported in the literature. Section 3 proposes the new reporting guidelines. Section 4 provides some conclusions.

2. PUBLISHED REPLICATIONS

As a starting point for the proposed guidelines, I performed a small literature review. This review focused on replications that were published in the International Symposium on Empirical Software Engineering and in *Empirical Software Engineering: An International Journal*, the main conference and journal of the empirical software engineering community. While there are replications published in other venues, I focused my review on these venues under the assumption that the replication papers published there would be the most complete and consistent because the empirical software engineering community is the most experienced at performing and publishing experiments and replications. Section 2.1 discusses the process of identifying the papers included in the review. Section 2.2 illustrates the different approaches these papers took in discussing the original study. Section 2.3 focuses on how the papers compare the results of the replication with the results of the original study. Finally, Section

Carver, J., "Towards Reporting Guidelines for Experimental Replications: A Proposal." Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER) [Held during ICSE 2010]. May 4, 2010. Cape Town, South Africa.