What Makes a Good Empirical Software Engineering Thesis?: Some Advice

Sira Vegas
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
svegas@fi.upm.es

***Abstract*— An empirical software engineering (ESE) PhD thesis has some special features, which makes it slightly different from a thesis any other different field. One of the differences between the two is the intensive use of empirical studies in an ESE dissertation. This talk starts by giving students advice on what makes a good ESE PhD thesis in the form of a list of do's and don'ts. The keynote later discusses what different empirical studies can be used (surveys, case studies and experiments). Finally, it focuses on one specific type of empirical study: controlled experiments. Experimentation is a risky business, and software engineering (SE) has some special features, leading to some experimentation issues being conceived of differently than in other disciplines. Some advice is given on how to analyse SE experiments.**

Keywords— empirical software engineering, PhD thesis.

I. What is an Empirical Software Engineering Thesis?

A few years ago, PhD theses did not use to include empirical validation. However, times are changing, and, nowadays, any PhD thesis must include at least some kind of empirical validation.

This does not mean, however, that any PhD thesis with an empirical component is an empirical software engineering (ESE) thesis. The key characteristic of ESE theses is that they have a major empirical component.

There are different types of ESE PhD theses. There is no formal typology, but, if we look at the ESE PhD theses written over the last 20 years, two main types stand out: 1) theses gathering knowledge about a specific topic by means of empirical studies, and 2) theses proposing methodological advances in ESE.

A. *Theses Gathering Knowledge about a Specific Topic*

Examples of such theses are (in chronological order):

- Seaman [11] conducts an empirical study whose goal is to characterize certain aspects of communication among members of a software development organization.
- Shull [13] runs a series of experiments to develop a body of knowledge on reading techniques for inspections.
- Thelin [15] reports a series of experiments on a reading technique called *usage-based reading*.
- Carver [1] studies the impact of an inspector's characteristics (background and experience) on his or her effectiveness in a software inspection.

The thesis may, in some cases, take in methodological aspects. For example, Seaman addresses the problem of how to analyse qualitative data. Shull tackles the problem of synthesizing the results of the different studies run. In other cases, techniques that have not been used in SE before are applied. For example, Carver uses grounded theory [5].

B. *Theses Proposing Methodological Advances in ESE*

Examples of such theses are (in chronological order):

- Daly [3] proposes a multi-method approach to empirical research, which, when integrated with the technique of replication, outputs more reliable and generalizable results.
- Ciolkowski [2] proposes an approach for the quantitative aggregation of evidence from controlled experiments in software engineering (SE).
- Jedlitschka [6] deals with the problem of reporting the results in SE experiments so that they are useful for software managers for decision making.
- Solari [14] addresses which contents a laboratory package for running SE experiments should have.
- Gómez [10] proposes a taxonomy for replications in SE. The taxonomy is used as a driver to plan the order in which replications should be run and how their results should be aggregated.

In all cases, empirical studies are run (or used) as a means to develop and validate the research performed.

II. Do's and Don'ts

Irrespective of the thesis type, some general guidelines can be established around three key issues that a PhD thesis should address: tackled problem, research method used and publication.

A. *Tackled Problem*

Regarding the *definition of the problem*:

- DON´T take it for granted that everybody is aware of

the problem.

- DO clearly specify what problem you are tackling.

Regarding the *scope of the problem*:

- DON´T try to solve a huge problem.
- DO define a problem with a scope that is reasonable for the time frame of a PhD thesis (your advisor will help with this).

Regarding the *importance of the problem:*

- DON´T think that you believing that the problem is important is enough.
- DO objectively assess the importance of the problem so that you can establish that the problem exists and that it is important. There are several ways to do this. One way is by citing other authors that state that the problem is important. Another is by citing numbers taken from a reliable source (for example, "the testing process needs improvement as the process in place fails to detect X% of the defects before the software is delivered to the users").

Regarding the fact that *the problem has not yet been solved*:

- DON´T take it for granted that people will trust you when you say that the problem has not yet been solved.
- DO demonstrate that your work contributes to the advancement of the state of the art/practice (systematic literature reviews might be helpful here).

B. Research Method

Regarding the *selection of the research method*:

- DON´T start working until you have sketched your research method. This will save you from wasting time.
- DO explain and properly justify the research method that you have chosen (remember that ESE PhD thesis have a very strong empirical component; therefore, your research method should be empirical).

Regarding the *research plan*:

- DON´T do uncontrolled research.
- DO draw up a research plan. This will help you to apply your method, and keep track of possible deviations in contents and time.

Regarding the evaluation/validation method:

- DON´T forget that you need to evaluate/validate your proposal. Different types of thesis require different types of evaluation/validation.
- DO choose the empirical study that best fits your research.

C. Publication

Regarding *dissemination of results*:

- DON´T postpone publication until you have finished your PhD thesis. Although this was the standard approach years ago, it does not work like that anymore. Some universities require you to have published at least a conference or journal paper before the defence of your thesis.
- DO try to publish results as early as possible (the state of the art could be a good choice). Of course, this does not mean that you should publish non-conclusive results.

Regarding *writing the dissertation*:

- DON´T think that publishing will be a waste of time that might be better spent on advancing in your research.
- DO consider other options for writing your thesis. Some universities accept PhD thesis formats other than the *traditional* dissertation. For example, each thesis chapter is styled as a paper. This will speed up the process of writing your thesis and will be an incentive for publishing.

III. Types of Empirical Studies

There are different types of empirical studies [16]: surveys, case studies, controlled experiments and quasi-experiments. All of them can be used in an ESE thesis:

- A *survey* is a method for collecting information from or about people to describe, compare or explain their knowledge, attitudes and behaviour [4].
- A *case study* is an empirical study that draws on multiple sources of evidence to investigate one instance (or a small number of instances) of a contemporary SE phenomenon within its real-life context, especially when the boundary between phenomenon and context cannot be clearly specified [9].
- A *controlled experiment* (or simply experiment) [7] is an investigation that establishes a particular set of circumstances (treatments) under a specified protocol - established and controlled by the investigator- to observe and evaluate implications of the resulting observations (dependent variables). SE works with *comparative experiments*, which implies: 1) the establishment of more than one treatment, and 2) responses resulting from the differing treatments are compared with one another [7]. The purpose of a controlled experiment is to identify causal inference.
- A *quasi-experiment* is an experiment where the assignment of treatments to experimental units (subjects) has not been randomized [12]. Assignment is made by means of self-selection (units choose treatment for themselves) or administrator selection (researchers decide which subject should get which treatment).

According to Pfleeger [8] and Wohlin *et al*. [16], several factors should be taken into consideration when deciding the type of empirical study to be used: 1) how much control the

experimenter has over the study; 2) the degree to which the researcher can decide which measures are to be collected; 3) the cost of the investigation and; 4) the easiness of replicating the investigation. **TABLE I** [16] shows how these factors vary with each empirical study.

TABLE I. FACTORS AFFECTING EMPIRICAL STUDIES.

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

IV. ISSUES WHEN ANALYZING EXPERIMENTS

Controlled experiments are very common in SE today. However, this is a challenging error-prone activity. Some common pitfalls that should be avoided are next discussed.

A. One- vs. Two-Tailed Tests

Using one-tailed tests implies predicting the direction of the effect. One-tailed tests are more powerful than two-tailed tests (we need a smaller test statistic to find a significant result). However, if the result of a one-tailed test is in the opposite direction to what you expected, you cannot reject the null hypothesis, and you will have to disregard the result.

B. Matching Data Analysis and Experimental Design

Data analysis is driven by the experimental design. Issues such as the scale used to measure the treatments and dependent variables, the number of factors and whether the experiment has a between- or within–subjects design, will determine the particular data analysis technique to be applied.

However, the choice of data analysis technique and/or statistical model is sometimes not straightforward. Complex designs may require the addition of some extra factors (and possibly interactions) to the statistical model. Take, for example, designs with blocking variables; the blocking variables and their interactions with treatments have to be added as factors to the statistical model. Another example are crossover designs; the order in which subjects apply treatments (sequences) and the times at which each treatment is applied (periods) have to be added to the analysis as factors.

C. What to Do when Test Assumptions Are Not Met

Parametric tests are more powerful than non-parametric tests and are capable of analysing several factors and their interactions. But the data do not always meet the parametric tests assumptions (typically normality and/or homogeneity of variances). However, data transformation and robust tests are an alternative to non-parametric tests.

D. Effect Size

Statistical significance measures whether the observed effect is the result of treatments or sampling error. It gives no indication of how big the difference in treatments is. For relatively large sample sizes, even very small differences in treatments may be statistically significant. If we want to know whether the differences between treatments are large enough to be of practical importance, we need a measure of effect size. There are different types of effect size measures.

E. Power Analysis.

A priori power analysis is used before the experiment is run to calculate the minimum sample size required for detecting an effect of a given size. Of course, a bigger sample size will be needed to detect small effects than medium or large effects.

Post-hoc power analysis determines the power of a given study assuming that the effect size of the sample is equal to the population. While the utility of a priori power analysis is universally accepted, the usefulness of post-hoc power analysis is controversial (it is a function of the statistical significance).

ACKNOWLEDGMENT

Research funded by the Spanish Ministry of Economy and Competitiveness research grant TIN2014-60490-P.

REFERENCES

[1] J. Carver. The Impact of Background and Experience on Software Inspections, PhD Thesis. Department of Computer Science, University of Maryland, Technical Report CS-TR-4476, April 2003.

[2] M. Ciolkowski. *An Approach for Quantitative Aggregation of Evidence from Controlled Experiments in Software Engineering*. PhD Theses in Experimental Software Engineering, Vol. 42. Fraunhofer Verlag, 2012.

[3] J.W. Daly. Replication and a Multi-Method Approach to Empirical Softwre Engineering Research, PhD Thesis. Department of Computer Science, University of Strathclyde, March 1996.

[4] A. Fink. *The Survey Handbook*, 2nd edition. SAGE, 2003.

[5] B.G. Glaser, A.L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter, 1967.

[6] A. Jedlitschka. *An Empirical Model of Software Managers. Information Needs for Software Engineering Technology Selection*. PhD Theses in Experimental Software Engineering, Vol. 28. Fraunhofer Verlag, 2009.

[7] R.O. Kuehl. Design of Experiments: Statistical Principles of Research Design and Analysis, 2nd edition. Brooks/Cole Cengage Learning, 2000.

[8] S.L. Pfleeger. Experimental design and analysis in software enginering part 1-5. ACM Sigsoft Software Engineering Notes, 19(4):16-20, 20(1):22-26, 20(2):14-16, 20(3):13-15, 20(5):14-17. 1994-1995.

[9] P. Runeson, M. Höst, A.W. Rainer, B. Regnell. *Case Study Research in Software Engineering. Guidelines and Exmples*. Wiley, 2012.

[10] O.S. Gómez, N. Juristo, S. Vegas. Understanding Replication of Experiments in Software Engineering: A Classification. *Information and Software Technology*, 56(8):1033–1048, 2014.

[11] C. Seaman. Organizational Issues in Software Development: An Empirical Study of Communication, PhD Thesis. Department of Computer Science, University of Maryland, Technical Report CS-TR-3726, 1996.

[12] W.R. Shadish, T.D. Cook, D.T. Campbell. *Experimental and Quasi-experimental Designs for Generalized Causal Inference, 2nd edition*. Cengage Learning, Inc. 2001.

[13] F.J. Shull. Developing Techniques for Using Software Documents: A Series of Empirical Studies. Ph.D. Thesis. Department of Computer Science, University of Maryland. AAI9921012, 1998.

[14] M. Solari. Identifying Experimental Incidents in Software Engineering Replications. In *Proceedings of the International Symposium on Empirical Software Engineering and Measurement (ESEM'13)*, pp. 213-222, 2013.

[15] T. Thelin. Empirical Evaluations of Usage-Based Reading and Fault Content Estimation for Software Inspections, PhD Thesis. Department of Communication Systems. Lund Institute of Technology, 2002.

[16] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén. *Experimentation in Software Engineering*. Springer, 2012.