Paper Review CS846, Jan. 30

Aggregating empirical evidence for more trustworthy decisions **and** *The truth, the whole truth, and nothing but the truth: A pragmatic guide to assessing empirical evaluation*

Wenhan Zhu (Cosmos)
w65zhu@uwaterloo.ca
University of Waterloo
Waterloo, Canada

ABSTRACT

This week's chapter from the book being reviewed here is **Aggregating empirical evidence for more trustworthy decisions** by *Donald Budgen* and **The true, the whole true, and nothing but the true: A pragmatic guide to assessing empirical evaluation** by *Blackburn et al.*

Budgen's chapter is mainly about what makes an evidence to be true evidence and how can be make decisions based on systematic reviews created from the evidence and how exactly can we do it.

Blackburn et al. talked about giving claims and evaluation of the claims and what can go wrong when doing so in the field of computer science in programming languages and system. They gave a very in depth description of claims and evaluation and how they work together to introduce new ideas to the field. They also proposed a vision to have more papers with either great evaluation or great novelty ideas to be considered by the community when submitting to conferences.

KEYWORDS

paper review

1 SUMMARY

Budgen starts the chapter by using an example of when having pizza at one night and then had a bad dream later that night does not make having pizza an evidence of having the bad dream. However, when we are dealing with evidences in empirical software engineering, it's not that clear that whether one thing is the causation of the other or what is the evidence of something happening. In empirical studies, we need to take information from large number of human participants in order to make sure when repeating the study, variations will be small. When designing experiments, we want to take control of the independent variable however, it is not always true that we can achieve that. During the process, we also need to make decisions, but these decisions are not prune to bias. So evidence-based paradigm and systematic reviews are introduced as the 'model' for evidence-based decision-making. There are 5 steps of the process.

(1) Transform the information of some intervention to a question that can be answered.
(2) Determine the best evidence of the question by ways that are systematic, objective and unbiased.
(3) Critically evaluate the evidence from different points such as validity, impact and applicability.
(4) Find experts in the domain and have them evaluate it.
(5) Evaluate the results and use to improve the preceding steps.

Steps 1-3 forms a systematic review. It is also considered as a secondary study since it aggregate results from other studies and does not involve the original participants. The original studies are primary studies.

Despite, in original usage in medicine, when a doctor can give suggestion of which medicine to use base on systematic review. In the context of software engineering its often hard to make a decision. *Budgen* suggests it due to in the nature of medicine, the comparison are similar but in software engineering both the approach and evaluation might be different which makes gathering information during the systematic review hard.

Budgen also mentioned just as many other methods in software engineering, systematic review is no 'silver bullet' solution.

Blackburn et al. gave a guide to assessing empirical evaluations. The paper talks about the process of empirical evaluation and categorized them by sins. In empirical evaluation, we have a claim and an evaluation of the claim.

Sins of reasoning happens when the scope of the claim does not align with the scope of the evaluation. If we consider the Venn diagram of claim and evaluation. We can denote the areas in three parts, A, B and C, where A is claim only, B is both claim and evaluation, and C is evaluation only. When A is empty and B, C are not, we have sin of ignorance where we evaluated the claim with data that was not originally in the claim. When C is empty and A, B are not, we suffer from sin of inappropriateness where we did not evaluate all of the claim. And when B is empty and A, C are not, the claim and the evaluation are disjoint, and it's the sin of inconsistency since we are sort of comparing apples and oranges. And when all are not empty, we suffer from both sin of ignorance and sin of inappropriateness.

Sins of exposition can happen when the claim or evaluation does not convey it's ideas. Sin of inscrutability happens when the claim is not described adequately. It may happen when the author neglect part of the claim or the claim is ambiguous when it comes to wording. Since the claim is synthesized by it creator it should always be possible to make it clear. The sin of irreproducibility is a sin that happens when the evaluation is inadequate. It can happen when some of the evaluation phase is omitted either because of the author not knowing it's important or have to be removed due to constrains of length of work. When sins of exposition happens,

people will not be able to figure out if the claim suffer from the sins of reasoning.

To avoid sin of inscrutability we can make our best effort to describe the claim as clear and as unambiguous as possible. To avoid sin of irreproducibility we should state all relevant control variables whether we believe they are related or not.

To avoid sin of ignorance and sin of inappropriateness, we should consider the methods we use in evaluation and make sure it's viable in our claims. Although it seems to be simple, in practice there are many factors that can go wrong. To avoid sin of inconsistency, we need to make sure that we are evaluating what we actually want to evaluation. The sins of reasoning all seems to be simple and obvious, however, due to the lack of knowledge in practice is much more harder than we think they are.

The authors suggested evaluations of claims in other fields such that repeating existing work and having a third party evaluation will help assessing empirical studies.

2 THOUGHTS

The chapter and the paper talked about how empirical studies should be looked at and how to avoid problems in this process. Different fields in science have developed different ways of dealing with empirical studies and software engineering borrow many ideas but they don't always works as the place where they borrowed.

When conducting an empirical study it is important to make sure we are not biased and both the evaluation and claim are as clear as possible and provide enough information so that the study can be repeated. During the process, the selection of method and the used data needs to fit and make sure they are telling the "true" story.

3 RATINGS

I would rate *Budgen*'s chapter 3.5/5. It described the process of systematic review but I would like to see more concrete examples of how it's applied.

I would rate *Blackburn et al.*'s work a 4.5/5. It clearly states the difficulties and mistakes we face when conducting empirical studies and gave example of each of them and suggests a framework that we can follow and avoid making these mistakes.