Empirical Software Engineering, 5, 309–312, 2000.

Hints for Reviewing Empirical Work in Software Engineering

WALTER F. TICHY tichy@ira.uka.de
University of Karlsruhe, Karlsruhe, Germany

Papers about empirical work in software engineering are still somewhat of a novelty and reviewers, especially those inexperienced with empirical work themselves, are often unsure whether a paper is good enough for publication. Conservative reviewers tend to err on the side of rejection, i.e., may sometimes reject a paper that, though not perfect, would advance our understanding or the empirical methods used. These hints are meant to help with judging empirical work and reduce some of the angst associated with accepting empirical papers.

1. Don't expect perfection.

Experiments are done in the real world and are therefore never perfect. Any empirical study, and especially a novel one, has flaws.

It is acceptable for a study to have threats to validity but the authors must mention them and take account of them in drawing conclusions and generalizing their results.

Sometimes, while a paper's results may not be strong in themselves, the approach to the research question may be novel and ingenious. Publication may later lead other researchers to take a similar approach.

However, do not hesitate to reject a paper if the results are seriously flawed or the methodological approach has no merit.

2. Don't expect a chapter out of a statistics textbook.

The examples in statistics textbooks are highly polished and carefully selected for illustrating statistical concepts. They are not representative of empirical work. Empirical data is always "dirty." Judging empirical work by the standards of statistics texts would lead to the suppression of many interesting results. Imperfections in the data or the data collection methods are tolerable, if they are unlikely to seriously affect the results.

If you find weaknesses in the statistical analysis, advise the authors how to correct them. Sometimes a more thorough analysis teases out additional issues that strengthen the paper.

Do not criticize authors for using unfamiliar statistical methods, such as resampling (aka Bootstrap). Do ask the authors to provide appropriate references or to explain a method, say in a footnote, side bar, or appendix.

Be critical about "cooked" data. A minimal amount of processing makes the results easier to interpret and more trustworthy than the output of a series of magical transforms.

Keep in mind that statistics is merely a tool, a servant in a broader inquiry.

3. Don't expect decisive answers.

Reviewers new to empirical work sometimes expect "the final answer" to a question from an experiment, as if experiments were as decisive as mathematical proofs. The reality of even the most rigorous approach to empirical work is that experiments normally constitute only a small step forward. By their very nature, experiments explore the relationships between a few variables only, while the real world is far more complex. Due to their limited scope, experiments merely gather evidence. Even when rejecting a hypothesis, a single experiment is often not decisive enough.

Rather than expecting "the final answer," ask whether the work contributes to our knowledge about a certain theory or set of questions. Small steps in the right direction will eventually get us there. If we cut off the small steps, then it is unlikely that we'll go the distance. Big leaps are rare.

4. Don't reject "obvious" results.

Some reviewers react to empirical results with "I could have told you that from the start" or "That's obvious." This is because reviewers are selected for their expertise in a particular area, in which they have lots of experience or about which they hold strong beliefs. For example, someone who collects design patterns might find any experiment testing their usefulness superfluous. He or she is already convinced! Keep in mind, though, that there may be skeptics. It is much easier to win over skeptics (i.e., seasoned practicioners) with data than with advocacy. Even if the last skeptic has been evangelized, we might all be wrong. And even if we all had the right idea all along, it may still be useful to check the magnitude of the effects of a certain programming technique, or observe the mechanisms causing a well-known effect.

All of these points argue for empirical studies that check beliefs and anecdotal evidence.

Although most of the assumptions underlying software tools and techniques have been insufficiently validated, this does not mean we should publish every little study about them. The research question underlying an empirical paper must be interesting in the sense that answering it would further our understanding of how software is produced. But don't simply reject a paper as "obvious" on the basis of personally held views or personal experience.

5. Don't be casual about asking authors to redo their experiment.

Reviewers who find flaws are quick to ask the authors to redo their experiment. However, rerunning an experiment may be way too expensive or time consuming. Redoing an experiment is not comparable with redoing a proof, rerunning a benchmark, fixing a bug, or refining a statistical analysis. The latter can often be done by the authors in a relatively short time. Experiments with human subjects, on the other hand, require the motivation and commitment of dozens of people, most of which will not gain from participating in the study. They may have to be trained for weeks or months before the experiment can start. A team of dedicated experimentalists is doing quite well if they can complete one or two experiments a year. Asking them to redo an experiment may well set them back by half to a full year, until the necessary constellation of subjects, preparation of subjects, and motivation is again possible. Worse yet, when using professionals from industry, an experimenter usually does not get a second chance.

If authors clearly explain the weaknesses of their experiment and take them into account in their conclusions, then don't ask them to repeat the study. Textbook-like perfection is too much to ask. Instead, ask whether something can be learned from the experiment as is, or whether the analysis of the data could be strengthened. Our science is served better by publishing good papers regularly than perfect ones every once in a while.

However, if the weaknesses of an experiment render it basically meaningless, then the best advice may well be to redo it, or to do a different experiment.

6. Don't dismiss a paper merely for using students as subjects.

Some reviewers consider an empirical study worthless if it hasn't been performed with professionals. They would like to see every study done with hundreds of programmers in a real project lasting two or three years, accompanied by a control group of the same size and duration. As ideal as this may sound, we live in a world where we have to make progress under less than ideal circumstances, not the least of which are lack of time and lack of willing subjects.

Situations where students are acceptable include the following:

a. Student subjects have been trained sufficiently well to perform the tasks asked of them. They must not be overwhelmed by the complexity of, or unfamiliarity with, the tasks or domain. Obviously, one can only study behavior that will occur.

b. Student subjects are used to establish a trend. Say a study compares two methods to see which is better. If one method has a clear relative advantage over the other with student subjects, then one can make the argument that there will be a difference in the same direction (although perhaps of a different magnitude) for professionals, provided the professionals use the methods in a similar fashion.

c. Student subjects are used to eliminate alternate hypotheses. Suppose an experiment with student subjects shows no clear difference between two methods. Unless there is evidence of radically different approaches by professionals, then it is highly unlikely that a noticeable effect will magically appear among professionals. It will also be nearly impossible to find professionals to participate in a follow-up experiment in this case. Therefore, allow negative results with student subjects to be published and thereby help the community discard wrong assumptions and move on.

d. Studies with students are a prerequisite for getting professionals to participate. It is hard to overemphasize this point: Experiments *must* be tested and debugged with students before running them with professionals. The experimental design and the trends found may be worth publishing, even if the follow-up experiment is the "real thing."

Experiments with psychology students have often been criticized for generalizing from students to the general population. University students are not representative of the general population with respect to a host of issues. Does the same criticism apply to experiments with CS students?

I think computer science students are much closer to the world of software professionals than psychology students are to the general population. In particular, CS graduate students

are so close to professional status that the differences are marginal. If anything, CS graduate students are technically more up to date than the "average" software developer who may not even have a degree in CS. The "professional," on the other hand, may be better prepared in the application area and may have learnt to deal with systems and organizations of larger scale than a student.

Studies have found that mere length of professional experience has little to do with competence. In other words, you can't use the argument that professionals with years of experience will necessarily solve a given problem better than appropriately prepared (graduate) students. If scale or application experience matters, then the story may be different.

7. Don't reject negative results.

Negative results, if trustworthy, are extremely important for narrowing down the search space. They eliminate useless hypotheses and thus reorient and speed up the search for better approaches. It may be appropriate to publish negative results as short papers.

8. Don't reject repetition of experiments.

A hallmark of empirical work is that a theory is provisionally accepted only if it has been tested in at least two experiments by independent groups. So, repeating experiments is a necessity. A repetition usually extends or modifies the original work and hence can broaden the results. Identical repetitions with identical results can be published as short papers. Repetitions become highly important, of course, if they come up with different results!

Experiments are difficult because they face reality and reality is often unpredictable and more complicated than anticipated. Reviewers need to understand that an experiment taking place in the real world cannot be controlled in the same way as a mathematical theorem or a software system. Empirical papers must be judged by standards that take into account the interface into unpredictable, uncooperative reality. But at the same time, this interaction with reality is what makes empirical work fascinating and rewarding.

Acknowledgments

I am grateful for helpful comments by Vic Basili, Warren Harrison, Anneliese von Mayrhauser, Norman Schneidewind, Elaine Weyuker, and Stu Zweben.