**Once is not enough:**

**Why we need replication**

Natalia Juristo

**1.- Motivating Example and Tips**

Imagine you land on an uncharted planet where no human has been before. For one reason or another, you are unable to leave the spaceship to explore the unknown, and you can only gather information about the planet through the spaceship windows. You look out and see a strange being: an alien of a shape, size and color that you could have never imagined. You are extremely surprised, and you carry on looking at it in order to unlock the mystery and form an idea of what sort of thing you have in your sights. Are you sure that what your eyes are seeing (and your brain is perceiving) through the window really is a true likeness of the outside world? What if the food you had to eat last night had gone off? What if you are seeing things, and there really is nothing out there? What would you do to make sure?

***Tip****: Empirical studies results should not be blindly trusted. Empirical data can also lie. How to be sure that what you observe exist?*

For ruling out the possibility that you are seeing things you ask one of your travelling companions to look out of the window as well. You will not tell her anything about what you have seen. You lead her to the window and ask her what she can see, taking care not to manipulate her in any way. She looks out of the window next to yours and cannot see anything at all. Is that so? You go round to the window where she is stationed and find that you cannot see the alien from there either, while she takes her place at your window. Aha! There is a light of some kind shining into this window, and it makes the alien invisible from here. Your travelling companion confirms that she, too, can see the alien from your window. It does exist then; it’s out there!

***Tip:*** *Site and researchers are variables that might induce different results of the same identical empirical study. Different researchers being able to reproduce your results in their labs, increase confidence in your empirical results. Is that all we need?*

“There is a red living being out there,” she says. Are you two sure enough that the alien is really red? Could it be that the glass in the window is so thick that it is deforming the alien’s profile? What if the glass is not perfectly transparent and is altering his color?

“Al is watching us!” your colleague shout astonished (by this time you have given the creature a pet name). “Watching us?” you ask with surprise, “I can’t see any eyes.” You and she each take it in turns to look out of the window, and you discuss and compare what you see. What she referred to as an eye looks to you like a blotch on its skin (if you can call the alien’s outer covering skin…). She decides to climb up to the observatory on top of the spaceship to study Al using the onboard telescope. Great! Thanks to this new instrument, she can distinguish details that were unappreciable by just looking through the windows. The information that she proffers from the telescope is crucial making out the eye-blotch controversy as well as some particulars that were unclear from the window.

***Tip:*** *Instruments matter. Even the best-designed empirical study interferes with the phenomenon at hand. It is of utmost importance that the same phenomenon is studied with different instruments in order to guarantee that the observation is independent of the instrument.*

**2.- Exploring the Unknown**

This imaginary scenario has a lot in common with empirical research work. We empirical software engineering researchers are every inch the spaceship’s crew on our way to an unexplored planet. We are on a voyage into the unknown: software development. Unfortunately, we cannot *penetrate* the unknown. We cannot travel to the place where the strings of software development are pulled; the place where the development variables causing the external behavior that we observe in software projects are cooked up. Software *backstage* where hidden variables are ruling development behavior is equally as impenetrable to our senses as gravity is. We human beings cannot perceive the gravitational forces that are at work behind the behaviors that we observe (for example, when we spill our coffee or when the Earth obediently moves in its orbit around the Sun). Likewise, we cannot directly observe the relationships between the variables causing the behaviors that we observe in software development. The only option open to us is to gather information about software development indirectly through empirical studies.

Any empirical study is a window onto the software development backstage. But the prospect from such window gives us only one view of the reality that we are scrutinizing. Unfortunately, the window is not open; it contains a piece of glass that has a bearing on what we see. Thus, we have to take the same precautions as our friends on the spaceship that just landed on the unexplored planet.

**3.- Types of Empirical Results**

If an experiment is not replicated, there is no way to distinguish whether results were produced by chance (the observed event occurred accidentally), results are artifactual (the event occurred because of the experimental configuration but does not exist in reality) or results conform to a pattern existing in reality. Different replication types help to clarify which of these three types of results an experiment yields.

When you see something shocking, striking and unexpected, what do you do to be sure? You look twice. To build a piece of reliable knowledge from an empirical study, you need to be sure that the observed results did not happen accidentally. So, do it twice! Repeat the study: same experimenters, same lab, same instrumentation, and same protocol. To rule out by chance results and get an estimation of the natural variation of the observation we need identical repetitions.

Still if the results repeated in your lab, we need to know if they can be generalized to other sites and researchers. The way to get such level of certainty in empirical findings is that other researchers in their labs replicate your study. To rule out local and researcher-dependent results we need identical replications.

One empirical study, no matter how well designed it is, might produce artifactual results. There exist a relationship between reality and the observation instrument, no one individual empirical study can yield definitive results, as the observation instrument itself (the study setting) may be affecting some aspects of the findings. Other designs, other protocols, other instrumentations will be able to confirm if results hold. To rule out artifactual results we need conceptual replications.

One type of empirical study offers one view of the phenomenon under observation. In order to make our evidence more reliable, we have to go a step further and observe the phenomenon using another type of instrument. Other types of empirical studies observing the same phenomenon provide new perspectives. Different types of empirical studies (experiments, observational studies, historical studies, case studies, surveys…) and empirical paradigms (qualitative and quantitative approaches) provide complementary views of the reality that we are studying, and we can piece together a more accurate picture of the development phenomenon under study by synthesizing their results.

**4.- Do’s and Don't’s**

What is the moral of this story? Results from one single empirical study are just a preliminary piece of information; we should even consider it just an anecdote. Our studies have to be repeated, replicated and triangulated in order to form a reliable idea of any SE phenomenon that we investigate. Building a reliable piece of knowledge out of empirical studies requires:

* Do same study
  + by same researchers in same site. This type of repetitions is needed to get away from fortuity and start walking towards evidence.
  + by other researchers in a different site. Other researchers replicate the study with the same protocol as the baseline study. This strategy confirms that the results are independent of the researcher, site and sample.
  + with different protocol. Studies (of the same type) should be performed using different setting or protocol (operationalize differently the variables, use other study design, other measurement process, etc.). This strategy guarantees that the results are independent of the instrument used.
* Do a different study with same goals. Alternative studies should be conducted. Observe the same reality through different type of studies provides new information that cannot be gathered from the baseline empirical study type.

Noticed that leaving out steps might lead to uncertain situations. If you move directly to step three and results you get are different (which there is high probability to happen), you will be unable to trace back the source of variations since there are so many. The new information and the one you had will not fit together and nothing new can be learnt. Through baby steps each study will contribute with pieces fitting together like parts of a puzzle and the bigger picture will emerge.

Just because once happened that you observed something emerging from a set of data (being them either produced by your study or borrowed from a repository) *don’t* trust it. Replication is the tool science has of being sure that something observed really exists. Do it again!

To know more:

O. Gómez, N. Juristo, S. Vegas. *Understanding Replication of Experiments in Software Engineering: A Classification*. **Information and Software Technology**. 56, 8. 1033-1048. 2014

N. Juristo, S. Vegas. *The Role of Non-Exact Replications in Software Engineering Experiments*. **Empirical Software Engineering**. 16, 3. 295–324. 2011