Population Characterization Comes Before Sample Selection

Oscar Dieste
Universidad Politécnica de Madrid
Madrid, Spain
odieste@fi.upm.es

Davide Fucci
Blekinge Institute of Technology
Blekinge, Sweden
davide.fucci@bth.se

Valentina Lenarduzzi
University of Oulu
Oulu, Finland
valentina.lenarduzzi@oulu.fi

Sira Vegas
Universidad Politécnica de Madrid
Madrid, Spain
svegas@fi.upm.es

ABSTRACT

Software Engineering (SE) experiments typically have small, hard to acquire sample sizes. Recruiting subjects is an issue for SE progress. However, we argue that characterization, i.e., understanding the population traits, is the main problem.

CCS CONCEPTS

• **General and reference → Empirical studies**.

KEYWORDS

Participant selection, recruitment, characterization.

ACM Reference Format:
Oscar Dieste, Davide Fucci, Valentina Lenarduzzi, and Sira Vegas. 2022. Population Characterization Comes Before Sample Selection. In *Proceedings of International Workshop on Recruitment of Participants in Empirical SE (RoPES 2022).* ACM, New York, NY, USA, 2 pages.

1 THE CONTEXT

Empirical SE provides pieces of evidence for decision-making. What happens in other disciplines? Let's pretend we snoop the call of anxious parents to an online medical service:

Parents — Our kids have a 39° fever. Can we give them Aspirin?
Doctor — There are risks. Please give them Paracetamol instead.

P — We guess it is COVID-19. Grandpas have been exposed. Can we give them Aspirin if their temperature rises?
D — It would be wiser if your parents came in for a check-up. Elders usually have circulatory issues that require consideration.

P — We understand. And can we have it?
D — In most cases, yes, unless you have some special condition.

P — I am pregnant.
D — Pregnancy is not a condition, but thank you for saying. Aspirin is not usually recommended during pregnancy. Please ask your doctor before using Aspirin.

RoPES 2022, May 17, 2022, Pittsburgh, PA, USA

Aspirin is one of the most used drugs in the world. Even so, it was not administered in the three cases above. The reasons (Reye syndrome, blood clotting, pregnancy) are widely known [4]. Medical experiments are monitored seeking unexpected treatment effects. Such knowledge (in addition, of course, to case and longitudinal studies) inform practice and enables decision making.

That conversation is not possible in SE today. Even if we had an unquestioned effective technology (such as the Aspirin in Medicine), we would not know how to advise a development team when considering adoption. Let's take Test-Driven Development (TDD) as an example. Few SE areas, if any, has better empirical evidence than TDD (more than 100 primaries and seven secondary studies). The conversation with a development team leader would be as follows:

Team leader — Our product has quality issues. Can we adopt TDD to improve quality?
Scholar — I think so. The scientific consensus says that software quality increases when professionals use TDD [5, 7, 11, 15].

TL — What do you exactly mean by "professional"? Our team is young. Most of them are recent graduates.
S — Scientific studies do not usually give information about experience years. They split subjects into professionals and students.

TL — That seems a risk in our case!
S — Quality does not increase when students apply TDD. If your subjects are recent graduates, TDD will not improve it.

TL — Some seniors in our team are students yet. But they have substantial work experience.
S — I cannot tell with certainty. Motivation has probably impacted experiment results [15]. Learning ability/skills [15] and or test case design knowledge [2] influence experiment results too.

2 THE PROBLEM

Recruiting large sample sizes is perceived as *the* problem by many SE researchers. And they are quite right [13]. In our past research, we have experienced the adverse effects of populations with ambiguous traits, small sample sizes, or high attrition rates.

But assembling large cohorts is not the *key* problem. Rafique and Mišić [11] has more than 700 subjects. TDD experiments published since have increased the total sample size above one thousand. Perhaps other technologies will not raise a similar interest, and experiments will not be conducted in such large numbers in industry and academia. But even so, we can use recruiting platforms

(LinkedIn, ExpiWell, CrowdFlower, Zooniverse, Qualtrics, MechanicalTurk, Prolific, etc.) where large sample sizes are easy to acquire.

The critical problem is *characterization*, not recruiting. We understand characterization as **identifying the population traits relevant to a research problem**, rather than focusing on the similarities and differences between sub-populations such as students and professionals. The dialog above shows our position. The kids vs. grown-ups dichotomy has implications for Aspirin administration (Reye syndrome affects children and teenagers). Similarly, students and professionals achieve different productivity levels when using TDD. However, Aspirin intake does not depend on users being kids but on the users' health issues. Similarly, TDD adoption depends on professionals' traits (skills, test case design knowledge, etc.).

3 OUR POSITION

RoPES' topics of interest includes *recruiting the "right" participants* (in addition to the *students vs. professionals* issue addressed above). Such a thing as "right" or "wrong" participants probably does not exist. Populations can be characterized in strict terms theoretically. But experiments do not work with populations (unless they are very small, like a company's staff), but samples containing usually rather heterogeneous individuals. **We cannot get rid of such heterogeneity. We need to find ways to deal with it**.

Recruiting platforms are not a solution because they do not provide minute detail about their subjects' characteristics. For instance, Prolific seems more suited to research studies than MTurk, due to better pool management features [9]. Prolific subjects can be filtered using 100+ personal characteristics, such as demographics, languages, education, technology competencies, etc. However, these characteristics are not very detailed. Having a B.Sc. degree can be used as a filter, but the topics are quite abstract—e.g., we can find computing and computer science, but not software engineering.

In contrast, **SE researchers frequently collect very specific information about participants' characteristics**. For instance, Rainer and Wohlin [12] proposed a framework for assessing *credibility* when recruiting field studies participants (e.g., case studies, surveys). They augment Falessi et al. [3] R^3 model—which characterizes participants according to their experience being real, relevant, and recent—by defining three possible roles (different from functional roles, such as *Engineer Lead*) and five characteristics. The concept of credibility and consequently (parts) of Rainer and Wohlin's framework applies to SE experiments.

We cannot blame the platforms for their coarse-grained information. MTurk, and other platforms, were not specifically designed for experimentation [9]. Mturk provides a workforce to perform human-centric tasks. Prolific seems oriented towards sciences, such as psychology, which rely on a broader, general population for their studies—i.e., possessing characteristics that are more common, less specialized. Recent studies point out that the Prolific's subjects (and other similar platforms) are roughly equivalent to the general population, exceeding it in some traits, such as social media use and creation of online content [14], and education level [10].

When specific populations are required, platforms enable prescreening participants based on survey responses or participation in previous studies. These pre-screenings are the researchers' responsibility. SE researchers cannot rely on these platforms for acquiring the "right" samples. **Sample selection precedes platform usage and requires proper population characterization.**

4 THE WAY AHEAD

We propose to **identify the *distance* between the current sample and the target population, and apply strategies to reduce it to a minimum** [6]. This strategy applies, in principle, to both SE research scenarios ("classical" academic/industry experiments and experiments using recruitment platforms) outlined above.

Two mechanisms could fulfill these requirements: 1) a procedure to characterize the population (in line with Bergersen *et al.* [1]) and 2) a procedure to evaluate the similarity among samples and samples vs. populations (similar to Nagappan *et al.* [8]). Accordingly, we plan to perform the following steps:

- Identify the relevant sample/population characteristics.
- Select suitable instruments to measure these characteristics.
- Measure the distance between sample and population.
- Define strategies that could be used to reduce the distance.

Moreover, we aim to compare different recruiting strategies. To goal is to help researchers select the recruitment strategy with the highest return-on-investment—e.g., the cheaper approach that minimizes the distance between sample and population.

REFERENCES

[1] Gunnar R. Bergersen, Dag I. K. Sjoberg, and Tore Dyba. 2014. Construction and Validation of an Instrument for Measuring Programming Skill. *IEEE Transactions on Software Engineering* 40, 12 (2014), 1163–1184.

[2] A. Causevic, D. Sundmark, and S. Punnekkat. 2011. Factors Limiting Industrial Adoption of Test Driven Development: A Systematic Review. In *2011 Fourth IEEE International Conference on Software Testing, Verification and Validation*. 337–346.

[3] Davide Falessi, Natalia Juristo, Claes Wohlin, Burak Turhan, Jürgen Münch, Andreas Jedlitschka, and Markku Oivo. 2018. Empirical software engineering experts on the use of students and professionals in experiments. *Empirical Software Engineering* 23, 1 (2018), 452–489.

[4] National Center for Biotechnology Information. [n. d.]. PubChem Compound Summary for CID 2244, Aspirin. https://pubchem.ncbi.nlm.nih.gov/compound/Aspirin. Retrieved January 21, 2022.

[5] S. Kollanus. 2010. Test-Driven Development - Still a Promising Approach?. In *2010 Seventh International Conference on the Quality of Information and Communications Technology*. 403–408.

[6] Valentina Lenarduzzi, Oscar Dieste, Davide Fucci, and Sira Vegas. 2021. Towards a Methodology for Participant Selection in Software Engineering Experiments. *International Symposium on Empirical Software Engineering and Measurement*.

[7] Hussan Munir, Misagh Moayyed, and Kai Petersen. 2014. Considering rigor and relevance when evaluating test driven development: A systematic review. *Information and Software Technology* 56, 4 (2014), 375 – 394.

[8] Meiyappan Nagappan, Thomas Zimmermann, and Christian Bird. 2013. Diversity in software engineering research. In *9th joint meeting on foundations of software engineering*. 466–476.

[9] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[10] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current directions in psychological science* 23, 3 (2014), 184–188.

[11] Yahya Rafique and Vojislav B Mišić. 2012. The effects of test-driven development on external quality and productivity: A meta-analysis. *IEEE Transactions on Software Engineering* 39, 6 (2012), 835–856.

[12] Austen Rainer and Claes Wohlin. 2021. Recruiting credible participants for field studies in software engineering research. arXiv:2112.14186 [cs.SE]

[13] William Shadish and et al. 2002. *Experimental and quasi-experimental designs for generalized cusal inference*. WADSWORTH CENGAGE Learning.

[14] Aaron Shaw and Eszter Hargittai. 2021. Do the Online Activities of Amazon Mechanical Turk Workers Mirror Those of the General Population? A Comparison of Two Survey Samples. *International Journal of Communication* 15, 0 (2021).

[15] Burak Turhan, Lucas Layman, Madeline Diep, Hakan Erdogmus, and Forrest Shull. 2010. How Effective is Test-Driven Development?