TOWARDS BUILDING A SOLID EMPIRICAL BODY OF KNOWLEDGE IN TESTING TECHNIQUES

N. Juristo, A.M. Moreno, S. Vegas
Facultad de Informática. Universidad Politécnica de Madrid
Campus de Montegancedo.
28660 Boadilla del Monte, Madrid, SPAIN
{natalia,ammorenosvegas}@fi.upm.es

ABSTRACT
Testing technique-related empirical studies have been performed for 25 years. We have managed to accumulate a fair number of experiments in this time, which might lead us to think that we now could have a sizeable empirically backed body of knowledge (BoK) on testing techniques. However, the experiments in this field have some flaws, and, consequently, the empirical BoK we have on testing techniques is far from solid. In this paper, we use the results of a survey that we did on empirical testing techniques studies to identify and discuss solutions that could lead to the formation of a solid empirical BoK. The solutions are related to two fundamental experimental issues: (1) the rigorousness of the experimental design and analysis, and (2) the need for a series of community-wide agreements to coordinate empirical research and assure that studies ratify and complement each other.

Categories and Subject Descriptors
D.3.3 [Programming Languages]: Language Constructs and Features – abstract data types, polymorphism, control structures. This is just an example,

General Terms
Experimentation, Verification.

Keywords
Keywords are your own designated keywords.

1. INTRODUCTION
In [11] we performed a review of experiments on testing techniques to identify what knowledge the results of these studies supplied for helping in testing techniques selection for software projects. From this review, we discovered that, even though there were a lot of empirical studies on testing techniques, the knowledge generated by these studies were quite poor.

Additionally, one of the findings of [11] was that there is a need for some, as yet unavailable, solid empirical knowledge on testing techniques to be able to gain a deeper understanding of testing techniques behaviour. Therefore, the results of the experiments on testing techniques are vitally important.

For these two reasons, we have subsequently extended the survey we conducted, trying to specifically identify what of all that we know about testing techniques has been empirically demonstrated. This work is described in [7] The conclusion that we drew from this survey was that the results of the experiments conducted to date on testing techniques are not mature enough to provide a solid empirical BoK on testing techniques.

This article aims to go a step further in this direction by analysing in detail what the reasons underlying the empirical immaturity of the current BoK on testing techniques are. We also propose a series of problem-solving approaches. For this purpose, this paper describes what problems we have detected that stand in the way of the formation of an empirical BoK on testing techniques. It also lists a series of guidelines that would be of assistance for solving the detected problems.

The reviewed empirical studies can be grouped into several subsets, taking into account which techniques they compare. Specifically, we have five groups of studies:

- Studies on the data flow testing techniques family. Includes Weyuker [12] and Bieman & Schultz [2].
- Studies on the mutation testing techniques family. Includes Offut & Lee [10], Offut et al. [9] and Wong & Mathur [13].
- Comparative studies between the control flow and data flow testing techniques families. Includes Frankl & Weiss [5], Hutchins et al. [6] and Frankl & Iakounenko [3].
- Comparative studies between the mutation and data flow testing techniques families. Includes Frankl et al. [4] and Wong & Mathur [13].
- Comparative studies between the functional and control flow testing techniques families. Includes Basili & Selby [1], Kambites & Lott [8] and Wood et al. [14].

The characteristics that we have analysed to decide how mature is the knowledge provided by an experiment, can be grouped around three aspect, which are:

- Rigour when running an experiment. Includes: design rigour, data analysis rigour and findings beyond mere data analysis.
- Correspondence level between the experiment and the real world. Includes use of: meaningful programs and faults, response variable of interest and realistic use of the technique.
- Establishment of a community-wide testing techniques experimentation strategy designed to ease the combination of
experiments. Includes: existence of laboratory packages that allow replication, chaining of experiments and methodological advancement in the experimentation sequence.

2. GUIDELINES FOR MATURING THE TESTING TECHNIQUE EMPIRICAL BOK

2.1 Experiment Rigour

One essential feature of an experiment is that it should be fully defined. If an experiment is not properly defined we run the risk of its results not being valid. Typically, experiments are affected by what are known as threats to validity. Depending on the threads that affect a given experiment, the applicability and validity of the results will be conditioned. When an experiment is not well defined, we will not be able to determine which threats to validity affect the experiment, and therefore, we will not be able to interpret their results.

Another essential feature of an experiment is that the experimental data should be rigorously analysed. The way to achieve a reliable analysis is using data analysis techniques. If no such techniques are used to analyse the results of an experiment, we run the risk of misinterpreting the results, leading to mistaken findings. By just looking at the results of the experiment, the causes of the variation could be attributed to the factors when there could be other reasons behind this difference.

Empirical knowledge is gained by means of a series of successive goals/experimentation/results cycles. This means that an experiment should not stop at the mere analysis of the data output, but should also establish high-level findings from these data that can be traced back to the hypotheses and objectives of the experiment. In this manner, we could generate new hypotheses to refine the knowledge gained. A single experiment will never be enough for a piece of knowledge to be well enough understood to become part of a solid BoK. Several trials will certainly be needed until it is well enough known. Table 1 shows that the lack of rigour during experiment design is a fairly widespread problem. The most common problems in the analysed empirical studies are omission of experimental designs and parameter descriptions. Few details are usually given about the factors and their levels, and only the response variables are explained thoroughly. We can also see that, generally, around half of the studies do not back up their findings with statistical analyses, and the authors merely conduct a rough analysis based on the subjective interpretation of data. Finally, Table 1 shows that drawing high-level conclusions that can be used to make recommendations and taken up in other experiments is an important shortcoming of the empirical studies.

2.2 Level of Correspondence between the Experiment and the Real World

Another essential feature of experiments is that they should be run using objects that are representative of reality. In the particular case of testing techniques, the programs used during the experiment, and the faults that they contain should be representative enough of reality for the results of the experiment to be meaningful. It is very true that, because of their features, controlled laboratory experiments, which are the most abundant in Table 1, are smaller in size than other types of empirical studies. This means that they are going to work on small programs. Even so, programs taken from programming books rather than real projects are often used. These programs contain few faults (often one or two) that have in many cases been entered artificially, that is, they are not faults that have really occurred in the program, but have been inserted afterwards. However, the results of the experiments would be more meaningful if both the programs used in the experiments and the faults they contain were closer to the real world.

Returning to the idea of an experiments being representative of reality, apart from using objects that are representative of reality in experiments, we should also consider that the actual running of the experiment should be representative of the real world. More specifically, the testing techniques should be applied or used as they are in the real world. The role of testing techniques during the testing process is for the subjects to apply them (on their own or with the aid of some sort of tool if available) to get a set of test cases that are then executed to find the defects of the program. The danger of running an experiment that does not resemble the real world closely enough is that we could get findings that cannot be extrapolated to this reality. Looking at Table 1, we find that only three experiments include real applications of the techniques under examination. This is, the empirical study includes a series of subjects who are responsible for applying the techniques in question. The other studies all automatically generate (there are no subjects involved) test cases, primarily at random, until they get a set that meets the constraints of the technique. In these cases, the empirical studies have overlooked a crucial concern, namely, the human factors.

Besides, we find that, unfortunately, not all the empirical studies focus on response variables that are of interest for practitioners. A clear example of this is the response variable percentage of test cases that detect at least one fault, examined in two studies. In these cases, practitioners and researchers do not view technique effectiveness in the same way. Practitioners usually prefer alternative metrics for measuring effectiveness, such as the number of defects a technique detects or ratio of detected faults to total faults.

2.3 Community-wide Strategy to Ease the Combinations of Experiments

One essential feature of an experiment is that it should be able to be repeated under exactly the same circumstances. The fact that an experiment yields results should not be taken as evidence enough for these results to be considered a universal truth. It is the repetition of this at different places under if not equal then similar conditions that leads us to have more and more confidence in the results. For an experiment to be replicated, it is necessary to keep track of the conditions under which the experiment was run. This is the only way of assuring that the replication will be able to be performed under the same conditions. If an experiment is not described enough, it will be difficult (and sometimes impossible) to replicate. This means that the study cannot be included in the empirical knowledge maturity chain, unfortunately making the study useless. In order to prevent the problem of replications,
<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>DATA FLOW TESTING</th>
<th>MUTATION TESTING</th>
<th>CONTROL FLOW VS. DATA FLOW</th>
<th>MUTATION VS. DATA FLOW</th>
<th>FUNCTIONAL VS. CONTROL FLOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental design rigour</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Data analysis rigour</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Findings beyond mere analysis</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Use of programs/faults representative of reality</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Response variables of interest to practitioners</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Real technique application environment is taken into account</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>There are no topics remaining to be looked at or confirmed</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Experiment chaining</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Methodological advancement in experimentation sequence</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The empirical study fully meets the characteristic
The empirical study meets partially the characteristic
The empirical study does not meet the characteristic
replication packages should be created. However, it is not very clear what a replication package should contain. There is a need for the community to establish an standard for what should a replication package contain. Table 1, shows that the lack of replication packages very widespread problem. Only the experiments run by Kamsties & Lott and Wood et al. have the corresponding replication package available.

We have also analysed the questions that remain to be looked at or confirmed in the reviewed empirical studies. It is interesting to note that all the results of the reviewed experiments need be confirmed. We find that, in all cases, the studies have questions awaiting confirmation, mainly because they are isolated studies, in which there are neither exact replications nor replications that vary the research hypotheses or replications that extend the theory. Therefore, it is essential for the Experimental Software Engineering (ESE) community to reach agreement on and establish a more or less organised working framework that sets out the families of empirical studies that are needed for testing techniques so that the BoK grows in an ordered and organised manner, ruling out the now uncontrolled growth.

Finally, experiments are not enough to consolidate a piece of knowledge. The control in the empirical studies needs to be gradually relaxed until the hypothesis is tested in the real world, outside the laboratory. The advancement in the pursuit of knowledge should follow some sort of order. It makes no sense to conduct a case study without having run several controlled experiments, as cases studies are more costly and it is only reasonable that there should be some confidence in the truthfulness of the original hypotheses. Additionally, several replications of one and the same experiment will need to be run to get an acceptable level of confidence in a hypothesis. Table 1 shows that the sequence in the advancement of the maturity of empirical knowledge as regards the study types conducted is fairly chaotic. Often, observational studies are conducted without having first run any more than one experiment or even straight off without having run any experiment at all beforehand. In this respect, it would also be important for the ESE community to reach agreements to better combine the studies to assure that all the experimental efforts are useful and play a role in the construction of an empirical testing techniques BoK.

3. CONCLUSIONS

In this article, we have presented a survey of the problems concerning the experiments now existing on testing techniques. These problems prevent the 13 examined studies from being analysed as a whole.

We presented the experiments examined, along with the criteria for examining each one. Subsequently, we identified the issues affecting each experiment, taking into account these criteria. Also we established how frequently an issue crops up (very often, fairly often, not very often). Finally, we presented a series of guidelines that could fix the flaws in the experiments on testing techniques to date.

The identified guidelines refer to two important aspects in ESE: on the one hand, rigorousness as regards the experiment (this would include experimental design and analysis) and, on the other, the need to reach a series of community-wide agreements to coordinate empirical research and assure that the results yielded by the studies are ratified and complemented, enabling progress to be made in the formation of a solid and rigorous empirical BoK for testing techniques.

4. REFERENCES